



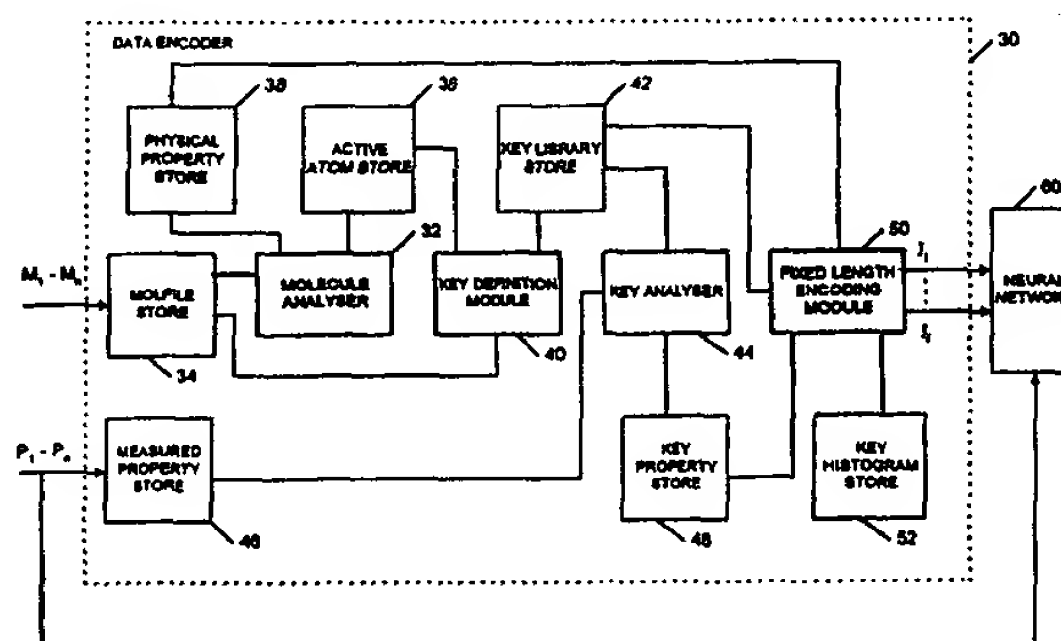
PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>6</sup> : <b>G06F 17/50</b>		<b>A1</b>	(11) International Publication Number: <b>WO 99/35599</b>
			(43) International Publication Date: 15 July 1999 (15.07.99)
(21) International Application Number: <b>PCT/GB99/00046</b> (22) International Filing Date: <b>7 January 1999 (07.01.99)</b> (30) Priority Data: 9800462.5      9 January 1998 (09.01.98)      GB (71) Applicant (for all designated States except US): EVERETT, Richard, Stephen, Hans [GB/GB]; Neural Computer Sciences, Unit 3, Lulworth Business Centre, Nutwood Way, Totton, Southampton SO40 3WW (GB). (72) Inventors; and (75) Inventors/Applicants (for US only): KETT, Brian, Laurence, Arthur [GB/GB]; 73 Compton Avenue, Lower Parkstone, Poole, Dorset BH14 8PX (GB). HARRIS, Richard, Alan [GB/GB]; 7A Forest View, Bargate, Southampton SO14 2BZ (GB). (74) Agents: BERESFORD, Keith, Denis, Lewis et al.; Beresford & Co, 2-5 Warwick Court, High Holborn, London WC1R 5DJ (GB).		(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).  Published With international search report.	

(54) Title: APPARATUS AND METHOD FOR USE IN THE MANUFACTURE OF CHEMICAL COMPOUNDS



(57) Abstract

A signal processing system is used in the manufacture of chemical compound to predict which molecules may have a required property for the compound. In the signal processing system, signals defining molecules in a training set and a measured property of each molecule are processed to produce fixed length signals encoding each molecule by (i) identifying for each molecule active atoms which may cause the molecule to react, and physical properties; (ii) defining each unique group which contacts a given number of the active atoms as a "key" in terms of the atoms and their relative positions; (iii) determining the contribution each key makes to the measured property of molecules containing the key; (iv) forming a histogram of the contributions of the keys in a molecule; and (v) encoding the molecule using the histogram values and the physical properties. The training data is used to train a neural network. A further molecule is encoded as before, but using the individual key contributions defined during training. The trained neural network is used to predict the property of the further molecule.

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

APPARATUS AND METHOD FOR USE IN THE MANUFACTURE OF  
CHEMICAL COMPOUNDS

The present invention relates to an apparatus and method  
5 for use as a tool in the manufacture of chemical  
compounds, for example medicaments. More particularly,  
the present invention relates to an encoding apparatus  
for use in a processing system which processes input  
signals to produce a signal predicting a property of a  
10 chemical molecule. The encoding system processes signals  
defining the molecule to produce encoded signals suitable  
for input to the processing system.

In the manufacture of new chemical compounds, it is often  
15 necessary to test many molecules to determine which ones  
have the required properties for the compound. For  
example, in the manufacture of a cancer drug, it is  
necessary to test molecules to determine the activity of  
each molecule against a cancer assay. Such tests are  
20 time consuming and expensive to perform.

Processing systems have been proposed which process input  
signals defining a molecule to predict a particular  
property of the molecule. These systems include neural  
25 networks and processing systems performing regression  
analysis.

Such systems, however, require the input signals defining the molecule to be of a fixed length (predetermined number of input bits). In neural networks, for example, an input neuron is provided for each element of the input signal. Accordingly, if the length of the input signal is unrestricted, then the neural network has to have an infinite number of input neurons in order to accommodate all possible inputs. In addition, it has been shown that the training of neural networks is less successful if the input signal is not distributed across all of the input neurons for all of the input molecules (for example if the input signals vary from a small length for some molecules which uses a small number of the input neurons, to a large size for other molecules which uses a large number of the input neurons).

"New Trends in Structure-Biodegradability Relationships" by Cambon and Devillers, Quant. Struct-Act. Relat. 12, 49-56 (1993) describes a neural network which is used to predict whether a molecule is weakly or highly biodegradable. Eleven structural features associated with persistent or degradable chemicals were identified from the literature and input molecules to the neural network were defined using an input signal of eleven binary digits, with each digit corresponding to one of the structural features and having a value of 1 if the structural feature was present in the input molecule and

0 if the structural feature was not present in the input molecule. A large number of structural features may be necessary to completely describe a large molecule. To decrease this number, the statistical method of correspondence factor analysis was performed on a matrix comprising the eleven Boolean values for each of the input molecules. The factor coordinates were then introduced as inputs to the neural network.

10 "A General QSAR Model for Predicting the Toxicity of Organic Chemicals to Luminescent Bacteria (Microtox® Test)" by Devillers, Bintein, Domine and Karcher in SAR and QSAR in Environmental Research, Vol 4, pages 29-38 describes the use of a backpropagation neural network to  
15 predict toxicity values of chemicals. An autocorrelation method was used to compute an autocorrelation vector encoding the hydrophobicity of the molecules and to compute an autocorrelation vector encoding the molecular refractivity. Stepwise regression analysis was then  
20 performed on the first ten components of these autocorrelation vectors to give nine autocorrelation components, on which principal components regression was carried out to produce a plurality of principal components. These principal components were then used  
25 as inputs to the neural network. The use of autocorrelation vectors to describe molecules is described further in "Autocorrelation of Properties

Distributed on Molecular Graphs" by Broto and Devillers  
in Practical Applications of Quantitative Structure-  
Activity Relationships (QSAR) in Environmental Chemistry  
and Toxicology (Karcher and Devillers Eds), Kluwer  
5 Academic Publishers, Dordrecht pages 105-127.

The encoding of the details of a molecule in a fixed  
length format results in a loss of information about the  
molecule, however. Accordingly, the subsequent accuracy  
10 of the system which uses the input signals to predict the  
properties of the molecules depends upon the information  
which is encoded and the way in which it is encoded.

The present invention has been made with the above  
15 problems in mind.

According to the present invention there is provided a  
signal processing apparatus or method in which a molecule  
is encoded on the basis of groups of atoms within the  
20 molecule.

The present invention provides a signal processing  
apparatus or method for use in training a processor such  
as a neural network, in which signals defining input  
25 molecules are processed to define groups of atoms within  
each molecule, a property of each group is calculated,  
and each molecule is encoded on the basis of the groups

within the molecule and the calculated properties of those groups.

The invention also provides a signal processing apparatus  
5 or method for use with a processor such as a trained  
neural network, in which signals defining an input  
molecule are processed to define groups of atoms within  
the molecule, and the molecule is encoded on the basis  
of the defined groups and a property of each group  
10 previously calculated during training.

The invention further comprises a process of  
manufacturing a chemical compound, such as a medicament,  
in which a property of a molecule is predicted on the  
15 basis of groups of atoms within the molecule, and the  
compound is made using a molecule predicted to have a  
suitable property for the compound (after testing of the  
molecule to confirm the predicted property if necessary).

20 The invention further comprises a compound containing a  
molecule predicted to have a property on the basis of  
groups of atoms within the molecule using a signal  
processing apparatus or method.

25 Embodiments of the invention will now be described by way  
of example only with reference to the accompanying  
drawings, in which:

Figure 1 shows the steps performed in an embodiment to manufacture a compound;

Figure 2 schematically shows the components of a  
5 processing apparatus used at step S2 in Figure 1;

Figure 3 shows a block diagram of the functional processing elements in the apparatus of Figure 2 used during analysis of known data to train the processor;  
10

Figure 4 shows the processing steps performed by the data encoder in Figure 3;

Figure 5 shows the structure for the example molecule  
15 glutaminyll;

Figure 6 shows the Molfile for glutaminyll;

Figure 7 schematically illustrates the information used  
20 to encode a plurality of atoms as a "key" for a molecule;

Figure 8 shows the keys for glutaminyll;

Figure 9 illustrates the key information stored within  
25 the key "library" store after the keys for one imaginary molecule have been identified;

Figures 10a, 10b, 10c and 10d show respectively the structure, Molfile, active atoms and keys for the molecule thymine;

5 Figures 11a, 11b, 11c and 11d show respectively the structure, Molfile, active atoms and keys for the molecule adenine;

10 Figures 12a, 12b, 12c and 12d show respectively the structure, Molfile, active atoms and keys for the molecule guanine;

Figure 13 illustrates the key information stored in the key library store after the keys for four imaginary  
15 molecules have been identified;

Figure 14 schematically illustrates the data stored in the key property store after a property value has been calculated for each individual key;

20

Figures 15a, 15b, 15c and 15d show key histograms for the example molecule 1, 2, 3 and 4 in Figure 13 respectively;

Figure 16 shows a block diagram of the functional  
25 processing elements in the processing apparatus of Figure 2 used during the analysis of a molecule to predict one or more of its properties;

Figure 17 shows the processing steps performed by the data encoder in Figure 15; and

Figure 18 shows the key histogram for the example  
5 molecule 1 in Figure 13 produced in a second embodiment.

Figure 1 shows the steps taken to manufacture a compound in an embodiment of the invention.

10 Referring to Figure 1, at step S2, signals defining molecules having known properties (for example as determined by experiment) are processed together with signals defining molecules with untested properties to predict the untested properties, (and hence determine  
15 which of the untested molecules may have the required properties to produce a compound with the desired characteristics).

Figure 2 shows a block diagram of the general arrangement  
20 of a signal processing apparatus used at step S2 to predict the properties of molecules. In the apparatus, there is provided a computer 2, which comprises a central processing unit (CPU) 4 connected to a memory 6 operable to store a program defining the operations to be  
25 performed by the CPU 4 and to store the signals processed by CPU 4.

Coupled to the memory 6 is a disk drive 8, which is operable to accept removable data storage media, such as a disk 10, and to transfer data stored thereon to the memory 6. Operating instructions for the central processing unit 4 may be input to the memory 6 from a removable data storage medium using the disk drive 8.

Data to be processed by the CPU 4 may also be input to the computer 2 from a removable data storage medium using disk drive 8. Alternatively, or in addition, data to be processed may be downloaded into memory 6 via a connection from a local or remote database which stores the data. The connection could, for example, be the Internet.

15

Coupled to an input port of CPU 4, there is a user-instruction input device 14, which may comprise, for example, a keyboard and/or a position-sensitive input device such as a mouse, a trackerball, etc.

20

Also coupled to the CPU 4 is a frame buffer 16, which comprises a memory unit arranged to store image data relating to at least one image generated by the central processing unit 4, for example by providing one (or several) memory location(s) for a pixel of the image. The value stored in the frame buffer for each pixel defines the colour or intensity of that pixel.

10

Coupled to the frame buffer 16 is a display unit 18 for displaying the image stored in the frame buffer 16 in a conventional manner. Also coupled to the frame buffer 16 is a video tape recorder (VTR) or other image recording device, such as a paper printer.

A mass storage device, such as a hard disk drive, having a high data storage capacity, is coupled to the memory 6 (typically via the CPU 4), and also to the frame buffer 16. The mass storage device 22 can receive data processed by the central processing unit 4 from the memory 6 or data from the frame buffer 16 to be displayed on display unit 18.

Data processed by CPU 4 and stored in memory 6 may also be recorded onto a removable data storage medium (such as a disk 10) using the disk drive 8, thereby enabling processed data to be exported from the machine. Processed data may also be exported by transmitting a signal conveying the data, for example, over a communication link (not shown), which could comprise the Internet.

CPU 4, memory 6, frame buffer 16, display unit 18 and mass storage device 22 may form part of a commercially available complete system, for example a conventional personal computer (PC).

Operating instructions for causing the computer 2 to perform as an embodiment of the invention can be supplied commercially in the form of programs stored on disk 10 or another data storage medium, or can be transmitted as  
5 a signal to computer 2, for example over a data link (not shown) so that the receiving computer 2 becomes reconfigured into an apparatus embodying the invention.

Processing is performed by computer 2 in two stages - a  
10 first stage to process signals defining known molecules and their known properties to train a neural network, and a second stage to process signals defining molecules with unknown properties to predict the properties using the trained neural network.

15

Figure 3 shows, as a block diagram, the functional elements within computer 2 used during the training stage to process signals defining known molecules and their known measured properties.

20

Referring to Figure 3, at a top level the functional elements comprise a data encoder 30, and a neural network 60. The components within the data encoder 30 will be described below with respect to the processing operations  
25 performed. The neural network 60 in this embodiment is a conventional backpropagation neural network with three layers, having 15 neurons in the input layer, 27 neurons

12

in the hidden layer and 1 neuron in the output layer. The instructions for causing computer 2 to be configured to have these functional elements may be input on a data storage device via disk drive 8, may be input over a communication link, for example from a remote source, or may be input directly using user-input device 14.

Signals defining a plurality of molecules,  $M_1-M_n$ , in terms of conventional "Molfiles" and signals defining a measured value,  $P_1-P_n$ , for a property of each respective molecule are input to data encoder 30. The molecules  $M_1-M_n$  and their properties  $P_1-P_n$  may be identified from the literature or from laboratory tests etc. Data encoder 30 processes the signals to produce a finite number of signals as inputs,  $I_1-I_{15}$ , to neural network 60 for each input compound M. These signals define an encoding of the input molecule M.

In this embodiment, data encoder 30 produces 15 signals ( $I_1-I_{15}$ ) to encode each molecule, as will now be described.

Figure 4 shows the processing steps performed in data encoder 30.

Referring to Figure 3 and Figure 4, at step S20, molecule

analyser 32 reads the Molfile defining the first input molecule  $M_1$  from the Molfile store 34 in which it was stored after input to data encoder 30. The Molfile specifies the atoms and their structural relationships  
5 within the molecule, and has the conventional format used in the ISIS system of MDL Information Systems Inc. By way of example, the Molfile for the molecule glutaminy1, whose structure is shown in Figure 5, is given in Figure 6.

10

Referring to Figure 6, the Molfile comprises the following conventional units as described in the ISIS documentation from MDL Information Systems Inc: a header block (containing background information such as users' initials, program name, date/time, dimensional codes,  
15 scaling factors, energy, and registry number), a counts line (which specifies the number of atoms, bonds and atom lists, the chiral flag setting and the connection table version), an atom block (which specifies the atomic symbol and any mass difference, charge, stereo chemistry,  
20 and associated hydrogens for each atom), and a bond block (which specifies the two atoms connected by each bond, the bond type, and any bond stereo chemistry and chain or ring properties).

25

Other units in the conventional connection table within the Molfile are not used in this embodiment.

Referring again to Figure 4, at step S22, molecule analyser 32 processes data defining the molecule read from the Molfile at step S20 to identify "active" atoms within the molecule, that is, atoms which are likely to  
5 react to contribute to the molecule having the measured property (for example when the molecule comes into contact with another molecule, or when the molecule is exposed to certain conditions).

10 In this embodiment, molecule analyser 32 identifies active atoms as atoms which satisfy one or more of the following conditions:

(1) If the atom is not hydrogen or carbon, it is  
15 identified as an active atom.

(2) If the atom is charged, it is identified as an active atom.

20 (3) Each aromatic ring in the molecule is considered to have a "virtual" atom at its centre which is active.

The active atoms identified at step S22 are stored in the  
25 active atom store 36.

Also at step S22, molecule analyser 32 processes the

15

information in the Molfile to determine the following physical properties for the input molecule:

- i) the number of aromatic rings in the molecule;
- 5 ii) the percentage of active atoms in the molecule;
- iii) the percentage of atoms which are carbon;
- 10 iv) the percentage of atoms which are oxygen;
- v) the percentage of atoms which are nitrogen.

By using percentage values for properties (ii) to (v) above, a number of predetermined size is obtained.

The physical properties calculated at step S22 are stored in the physical property store 38.

20 Referring, by way of example, to Figure 5 which shows the structure of glutaminy1, the processing performed by CPU 4 at step S22 would identify oxygen atoms 70, 72 and nitrogen atoms 74, 76 as active atoms because they are not hydrogen or carbon atoms. No atom in the glutaminy1  
25 molecule is charged and, therefore, no additional active atoms are identified using test (2) above. Similarly, there are no aromatic rings to produce "virtual" active

16

atoms (test (3) above). The processing performed at step S22 by CPU 4 would also identify the following physical properties for the glutaminy molecule:

5     -     aromatic ring count = 0

-     percentage of active atoms =  $4/19 = 21.05\%$

-     percentage of carbon atoms =  $5/19 = 26.32\%$

10

-     percentage of oxygen atoms =  $2/19 = 10.53\%$

-     percentage of nitrogen atoms =  $2/19 = 10.53\%$

15     Referring again to Figure 4, at step S24, key definition module 40 reads the active atoms identified at step S22 from the active atom store 36 and the Molfile from the Molfile store 34, identifies each unique group which comprises a predetermined number (in this embodiment 3)  
20     of the active atoms and encodes each group by defining the atoms in the group and their relative positions. This encoded information is referred to as a "key".

Figure 7 schematically illustrates the information  
25     encoded in a key.

Referring to Figure 7, each of the three atoms in a group

is encoded using its atomic number, and the relative positions of the atoms are encoded using the distance between each pair of atoms (in this embodiment this is defined as the smallest number of bonds which must be traversed within the molecule from one atom to the other) and the degree of freedom of the bonds between each pair of atoms (in this embodiment this is determined by adding the degree of freedom of each individual bond considered in the distance measurement for the atoms, with the degree of freedom of a bond being defined as 1 if it is a single bond (representing the ability of the bond to move) and 0 if the bond is a double bond, a triple bond, or if it is within a ring structure).

The key encoding performed in this embodiment enables compositional and topological data describing a molecule to be encoded in a fixed length signal. Also, the encoded information is independent of the conformational states of the molecule.

20

Figure 8 shows the keys for the molecule glutaminyll.

Referring to Figure 8, the first key listed comprises the numbers 778552541, which are derived as follows. The atoms in the first key comprise nitrogen atoms 74 and 76 and oxygen atom 70. These atoms have atomic numbers of 7, 7 and 8 respectively, which form the first three

numbers in the key.

The minimum number of bonds in the molecule between nitrogen atom 74 and nitrogen atom 76 is 5, and since  
5 each of these bonds is a single bond, each has a degree of freedom of 1 so that the total degree of freedom of the bonds between nitrogen atoms 74 and 76 is 5. The bond distance 5 forms the fourth number in the key and the degree of freedom number 5 forms the seventh number  
10 in the key. Similarly, the minimum number of bonds between oxygen atom 70 and nitrogen atom 76 is 5 (the fifth number in the key) but, since one of these bonds is a double bond, which is assigned a degree of freedom of 0, the total degree of freedom for the bonds is 4 (the  
15 eighth number in the key). The number of bonds between nitrogen atom 74 and oxygen atom 70 is 2, (the sixth number in the key) having a total degree of freedom of 1 (the ninth number in the key).

20 Although in Figure 8 the atomic numbers, network distances, and degrees of freedom are shown in a specific order, any order can be used, so that a given key can be written in a number of different ways. For example, the first key, 778552541, could equally be defined as  
25 877255145, for example.

The total number of keys for the glutaminy1 molecule is

four, as set out in Figure 8, since it is possible to define four unique groups which each contain three active atoms.

- 5 Referring again to Figure 4, at step S26, key definition module 40 stores the keys defined at step S24 in a key "library" within key library store 42.

Figure 9 schematically illustrates the storage of this  
10 information for an imaginary molecule (molecule 1) having four imaginary keys in the storage library.

In this embodiment, information is stored identifying each key and the molecule in which that key may be found.

15

- Referring again to Figure 4, at step S28, CPU 4 determines whether there is another molecule in the input training set. Steps S20 to S28 are repeated until all molecules in the training set have been processed in the  
20 manner described above to define and store their keys.

Figures 10a, 10b, 10c and 10d show respectively, by way of further example, the structure of the molecule thymine, its Molfile, its active atoms, and its keys.

- 25 Referring to Figure 10d, the degree of freedom of the bonds between each pair of atoms in each key is 0 because all of the active atoms are either in an aromatic ring

or connected to the aromatic ring with a double bond.

Figures 11a, 11b, 11c and 11d show respectively the structure, Molfile, active atoms and keys for the molecule adenine. Referring to Figure 11c, the virtual atom (marked \*) at the centre of the aromatic ring is defined as an active atom. Referring to Figure 11d, it will be seen that the atomic number of the virtual atom is defined as 0 in each key containing this atom, and that the network distance between any of the "real" active atoms and the virtual active atom is the minimum number of bonds which connects the real active atom to the aromatic ring which has the virtual atom at its centre. Figure 11d also shows that the key 770321000 (marked \*) appears twice in the molecule adenine.

Figures 12a, 12b, 12c and 12d show respectively the structure, Molfile, active atoms and keys for the molecule guanine. Referring to Figure 12d, the keys 777325011 and 777523110 are identical (as noted above, the numbers in a given key can be defined in different orders), and therefore this key occurs twice in the molecule guanine.

Figure 13 shows the information stored in key library store 42 after step S26 (Figure 4) has been repeated for all molecules in the training set. Figure 13 illustrates

the information stored for four imaginary molecules (not the real molecules described by way of example above) having ten imaginary keys. The information stored in key library store 42 defines a superset of the keys in the molecules (that is, each key in every molecule), and the molecules in which each key is found.

Referring again to Figure 4, at step S30, key analyser 44 reads the molecular properties  $P_1-P_n$  from a measured property store 46 in which they were stored after input to data encoder 30.

At step S32, key analyser 44 uses the key information stored in the key library store 42 at step S26 and the molecular properties read at step S30 to define a value representing the contribution each key makes to the property in a molecule.

Key analyser 44 determines the property for each key in dependence upon the molecules in the training set in which the key is found and the values of the property for each of those molecules. In this embodiment, key analyser 44 determines the contribution a given key makes to the property value  $P$  of each molecule in which the key is found as follows:

22

$$\begin{array}{lcl}
 \text{Contribution of key} & & \text{number of times key appears} \\
 \text{to property value } P_i & = & \text{in molecule } M_i \\
 \text{of molecule } M_i & & \hline
 5 & & \text{total number of molecules} \\
 & & \text{in which key is found} \\
 & & \dots\dots(1)
 \end{array}$$

Key analyser 44 then stores for each key the calculated contribution in key property store 48.

10

To illustrate the processing performed in this embodiment at step S32, an example will be given in which the property is, for example, the activity of each molecule against a predetermined cancer assay, having a value between 0.0 and 1.0, and in which, molecule 1 shown in Figure 13 has an activity ( $P_1$ ) of 0.5, molecule 2 has an activity ( $P_2$ ) of 0.4, molecule 3 has an activity ( $P_3$ ) of 0.8 and molecule 4 has an activity ( $P_4$ ) of 0.2.

Figure 14 shows the key information, including the calculated properties for each key in this illustrative example, stored in key property store 48 at step S32.

Referring to Figure 14, key 1 appears in two molecules, namely molecule 1 and molecule 3. Therefore, the contribution that key 1 makes to an activity level of 0.5 (this being the activity level of molecule 1) is  $\frac{1}{2}$  since key 1 appears once in molecule 1 and twice in all molecules. Similarly, the contribution that key 1 makes to an activity level of 0.8 (this being the activity

level of molecule 3) is also  $\frac{1}{4}$ . Key 1 does not contribute anything to an activity level of 0.4 (activity level of molecule 2) or 0.2 (the activity level of molecule 4) since it does not occur in these molecules.

5

Key 3 appears once in each of molecules 1, 2 and 3. Accordingly, the calculated property for key 3 is  $\frac{1}{3}$  at activity level 0.5,  $\frac{1}{3}$  at activity level 0.4, and  $\frac{1}{3}$  at activity level 0.8.

10

Key 6 appears once in molecule 2, twice in molecule 3 and once in molecule 4. The calculated property for key 6 is, therefore,  $\frac{1}{4}$  at activity level 0.4,  $\frac{1}{2}$  at activity level 0.8 (since key 6 appears twice in molecule 3) and  $\frac{1}{4}$  at activity level 0.2.

15

Referring again to Figure 4, at step S34, fixed length encoding molecule 50 uses the key properties defined at step S32 and the physical properties calculated at step S22 to encode each molecule in the training set in a fixed length format. In this embodiment, the encoding at step S34 is performed for each molecule by, firstly, reading the keys from the key library store 42 which are in the molecule (previously stored at step S26) and the property calculated for each of those keys from the key property store 48 (stored at step S32), calculating values comprising the sum of each of the individual key

20

25

properties, and, in effect, storing the calculated values in a histogram having a predetermined number of bins in key histogram store 52, and secondly, using the histogram numbers together with the numbers from the physical property store 38 calculated at step S22 for the physical  
5 properties as the encoding for the molecule. This processing therefore encodes arbitrary molecules of unknown size (unknown number of constituent atoms) with a predetermined number of numbers (which are defined  
10 using a predetermined number of bits in a digital signal).

Referring again to the illustrative example in Figure 14, for molecule 1, fixed length encoding module 50 adds the  
15 contribution that the keys in molecule 1 (that is, keys 1, 2, 3 and 4) make to activity level 0.5 and stores the total (1.83) in a histogram for the bin 0.5. Similarly, fixed length encoding module 50 adds the respective contributions that keys 1 to 4 make to activity level 0.4  
20 and stores the total (0.33) in the bin for 0.4 in the histogram, adds the respective contributions that keys 1 to 4 make to the activity level of 0.8 and stores the total (1.33) in the bin for 0.8 in the histogram, and adds the respective contributions that keys 1 to 4 make  
25 to the activity level of 0.2 and stores the total (0.5) in the bin for 0.2 in the histogram.

Figure 15a shows the histogram formed as described above for molecule 1. In this embodiment, CPU 4 provides 10 bins in the histogram, and therefore defines and stores ten histogram numbers for each molecule, the numbers for molecule 1 being 0, 0.5, 0, 0.33, 1.83, 0, 0, 1.33, 0, 0.

Fixed length encoding module 50 then uses the physical properties of the molecule previously calculated at step S22 to produce signals defining a fixed length encoding of the molecule. More particularly, CPU 4 uses the 10 values in the histogram for the molecule together with the values of the 5 physical properties calculated at step S22 to produce a signal having 15 values comprising an encoded format of the molecule.

Figures 15b, 15c and 15d show respectively the histograms formed at step S34 by fixed length encoding module 50 for molecules 2, 3 and 4 in the illustrative example of Figure 14. Again, fixed length encoding module 50 produces signals defining 15 values for each molecule comprising a respective value in each of the ten buckets of the histogram together with the respective value of each of the five physical properties previously calculated at step S22, thereby encoding each molecule with a fixed length format (15 numbers which can be defined using a predetermined number of bits).

Referring again to Figure 3, after performing the processing operations in data encoder 30 described above with respect to Figure 4, CPU 4 inputs each respective one of the 15 numbers  $I_1-I_{15}$  encoding a given input molecule M to a respective one of the input neurons of neural network 60. At the same time, CPU 4 applies a signal defining the measured property P of the molecule to the single output neuron of the neural network 60. This is done for each of the molecules  $M_1-M_n$  and their measured properties  $P_1-P_n$  in the training set to train the neural network in a conventional manner. This produces a trained neural network, which can then be used to predict the properties of molecules which are not in the training set.

15

Having trained the neural network 60, the first stage of processing is complete.

Figure 16 shows the functional processing elements in computer 2 used in the second stage of processing to predict the property  $P_{n+1}$  of a molecule  $M_{n+1}$  which was not in the training set. The elements are the same as those shown in, and described above with respect to, Figure 3, with the exception that key analyser 44 and measured property store 46 are not used during prediction (and hence are not shown in Figure 16) and key library store

42 is replaced by key store 54.

Referring to Figure 16, the Molfile defining the compound  $M_{n+1}$  to be processed is input to the data encoder 30,  
5 which now contains key property data stored in key property store 48 at step S32 (Figure 4) during training.

Figure 17 shows the processing operations performed by CPU 4 within the data encoder 30 during the second stage  
10 of processing.

Referring to Figure 16 and Figure 17, at step S100, molecule analyser 32 reads the input Molfile from Molfile store 34, and at step S102 identifies and stores the  
15 active atoms and physical properties of the input molecule in the same way that these were identified and stored at step S22 described above.

At step S104, key definition module 40 defines the keys  
20 within the input molecule in the same way that the keys were defined at step S24. Key definition module 40 stores the defined keys in key store 54.

At step S106, fixed length encoding module 50 reads the  
25 key properties previously defined on the basis of the training data at step S32 from the key property store 48.

At step S108, fixed length encoding module 50 encodes the input molecule using the keys in the input molecule  $M_{n+1}$  stored in key store 54 at step S104, the key properties previously stored in key property store 48 at step S32 during training, and the physical properties of the input molecule  $M_{n+1}$  stored in physical property store 38 at step S102. This encoding is performed in the same way that each molecule was encoded at step S34.

10 If the input molecule contains a key which is not one for which key properties are stored in key property store 48 (that is, the input molecule  $M_{n+1}$  contains a key which was not present in any of the molecules  $M_1-M_n$  in the training set) then, in this embodiment, the key is ignored by

15 fixed length encoding module 50.

Referring again to Figure 16, the 15 numbers  $I_1-I_{15}$  encoding the input molecule are input to the trained neural network 60, which outputs in response thereto the predicted value  $P_{n+1}$  for the property of the input molecule. Thus, referring to the example of the imaginary molecules described above, neural network 60 outputs a predicted value between 0 and 1 for the activity level of the input molecule against the

20

25 predetermined cancer assay.

After all input molecules which are not in the training set have been processed as described above, step S2 (Figure 1) is complete.

- 5 Referring again to Figure 1, at step S4, molecules predicted at step S2 to have a favourable property value (for example an activity level against a particular cancer assay which is greater than predetermined value) are synthesised and tested in a conventional manner in  
10 the laboratory, clinical trials etc.

At step S6, a compound is manufactured containing one or more of the molecules tested at step S4 which was found to have the required properties. Of course, the compound  
15 may include other molecules, for example as carriers etc.

The embodiment described above may be modified in a number of ways.

- 20 In the embodiment above, data encoder 30 and neural network 60 are provided in the same computer 2. Similarly, after training, prediction of the molecules' properties is carried out using the data encoder and trained neural network in the same computer.  
25 Alternatively, data defining the data encoder 30 with key property data stored in key property store 48 may be transferred to train a neural network in a different

computer. Similarly, data defining the data encoder 30 with the key property data stored in key property store 48 and/or data defining the trained neural network 60 may be transferred to a different computer in order to  
5 predict the properties of molecules not in the training set. As well, during prediction of properties (Figure 16) a data encoder without key data stored in key property store 48 may be used to identify active atoms and define keys (steps S100, S102 and S104 in Figure 17),  
10 and key properties may be read (step S106) from a key property store held elsewhere, such as a remote database, to enable encoding to be performed (step S108).

During training, if the number of input molecules in the  
15 training data which do not possess a particular property (for example they are inactive against a predetermined cancer assay) greatly outnumber the number of input molecules which possess the property, the neural network can become biased by the training. This is because most  
20 techniques will tend to err on the side of classifying an active compound as inactive because the cumulative error of misclassification of active compounds (false negatives) contributes little error when compared to the much larger accurate classification of inaccurate  
25 compounds (true negatives). This problem can be addressed by modifying the training algorithm or by using unbiased subsets of the training data. For example, in

one approach a training algorithm may be used which seeks to minimise the maximum error, such as the minimax algorithms discussed in "Neural Networks for Pattern Recognition" by C.M. Bishop, Oxford University Press, 5 1995, ISBN 0198538642.

In the embodiment above, at steps S22 and S102, molecule analyser 32 identifies active atoms by determining each atom which (i) is not hydrogen or carbon, (ii) is 10 charged, or (iii) is a virtual atom at the centre of an aromatic ring. Different conditions may be used instead of, or in addition to, these conditions to identify active atoms. For example, the following conditions could be used:

15

- the atom is any atom within an aromatic ring;
- the atom is carbon with at least one double bond to an another atom;
- 20 - the atom is not hydrogen and is bonded to at least one atom deemed active by any other condition.

The identification of active atoms at steps S22 and S102 25 could be omitted, and keys could be defined at steps S24 and S104 using all atoms in a molecule. However, this would significantly increase the number of keys generated

32

since, if keys are identified from among N atoms and the number of atoms in each key is n, then the number of unique keys is:

$$\frac{N!}{n! (N-n)!}$$

5

The physical properties identified by molecule analyser 32 at steps S22 and S102 may be different to those described in the embodiment above. For example, physical properties which may be used in addition to, or instead of some or all, of the physical properties described in the embodiment above include boiling point, melting point, freezing point, proportion of hydrogen, and whether the molecule is chiral.

15 In addition, some or all of the physical properties may be omitted. (If all of the physical properties are omitted, each molecule would be encoded using the numbers from the histogram produced at step S34 or step S108.)

20 In the embodiment above, at steps S24 and S104, key definition module 40 defines the atoms in a key on the basis of atomic number. In addition, or instead, mass number, valency and/or charge may be used. Also, in addition to, or instead of, encoding the topological arrangement of the atoms using the number of bonds between atoms in the key and the degree of freedom of the bonds, the absolute or relative distances between atoms

in the key and/or the angles between atoms in the key may be used. These measures, however, suffer from the problem that molecules are not static objects and, as they alter shape (conform), the distance and angle  
5 between atoms changes. This can be overcome to some extent by considering the minimal energy configuration (conformer).

Instead of defining keys on the basis of three atoms,  
10 four or more atoms may be used. This enables chirality to be modelled.

At steps S32/S34 and S108, different techniques can be used to encode a molecule. For example, at step S32, key  
15 analyser 44 may calculate an average property value for each key. Thus, referring to Figure 14 by way of example, the average property value for key 1 would be  $\frac{1}{2}(0.5 + 0.8)$ , since key 1 is in molecule 1 which has an activity level of 0.5 and molecule 3 which has an  
20 activity level of 0.8. Similarly, the average activities for keys 2, 3 and 4 would be respectively  $\frac{1}{2}(0.5 + 0.2)$ ,  $\frac{1}{3}(0.5 + 0.4 + 0.8)$  and  $\frac{1}{2}(0.5 + 0.8)$ . The average activity for key 6 would be  $\frac{1}{4}(0.4 + 0.8 + 0.8 + 0.2)$ , the activity level 0.8 being considered twice since key 6  
25 appears twice in molecule 3. At step S34, fixed length encoding module 50 would then calculate a histogram of the average activities for each key in the molecule being

encoded. By way of example, Figure 18 shows the key histogram produced for molecule 1 in the example of Figures 13 and 14 using this modified form of processing.

- 5 By way of further example, in the embodiment above at step S32, key analyser 44 may calculate the contribution of each key to a property value  $P_i$  of molecule  $M_i$  using equation (1) given in the embodiment above but without dividing by the number of molecules in which the key is  
10 found.

Further examples of the way in which the molecule may be encoded using the key information include:

- 15 - An additional neural network may be used to create a model which maps each key to the properties of the molecules in which the key is found. The creation of such a model would seek to minimise the residual error of the predicted activities. Once  
20 created, the model could be used as a measure of key activity for the construction of a histogram.
- A vector could be constructed with an entry for each key. Each element of the vector could be set  
25 to the number of times its associated key appears within the molecule in question.

In the embodiment above, a back propagation neural network is used. However, different types of neural networks may be used. For example, a radial basis function network or a multi-layer perceptron as described  
5 in "Neural Networks for Pattern Recognition" by C.M. Bishop, Oxford University Press, 1995, ISBN 0198538642 may be used. A Kohonen network as described in "Self-Organisation and Associative Memory 2nd Edition" by Kohonen, Springer Verlag, 1988, ISBN 038718140 may be  
10 used.

Neural network 60 may be replaced by a functional element performing conventional linear regression. Neural network 60 may also be replaced by a functional element  
15 comprising a generalised additive model, for example as described in "Generalised Additive Models" by T.J. Hastie and R.J. Tibshirani in Monographs on Statistical and Applied Probability, No. 43, Chapman & Hall. Neural network 60 may also be replaced by a functional element  
20 performing projection pursuit regression, for example as described in the Journal of the American Statistical Association, vol 76, No. 376, pages 817-823 by J. Friedman & W. Stutzle, 1981.

- 36 -

CLAIMS

1. A signal processing system, comprising:
  - (a) first encoding means, comprising:
    - 5 means for processing input signals defining a plurality of molecules to define groups of atoms within each molecule;
    - means for processing the defined groups and input signals defining a property of each of the plurality of
    - 10 molecules to define a characteristic of each group in dependence upon the property of molecules containing the group; and
    - means for generating signals defining each of the plurality of molecules in dependence upon the defined
    - 15 characteristics;
  - (b) second encoding means, comprising:
    - means for processing input signals defining a further molecule to define groups of atoms within the further molecule; and
    - 20 means for generating signals defining the further molecule using the defined groups in the further molecule and the characteristics of the groups defined by the first encoding means; and
  - (c) predicting means for generating a signal
  - 25 defining a predicted property of the further molecule by processing the signals generated by the first encoding means defining each of the plurality of molecules,

- 37 -

signals defining the property of each of the plurality of molecules, and the signals generated by the second encoding means defining the further molecule.

- 5     2.     A system according to claim 1, wherein the means for processing input signals in the first encoding means and the means for processing input signals in the second encoding means are each arranged to define groups containing a predetermined number of atoms.

10

3.     A system according to claim 2, wherein the means for processing input signals in the first encoding means and the means for processing input signals in the second encoding means are each arranged to define all unique  
15     groups which contain the predetermined number of atoms.

4.     A system according to any preceding claim, wherein the means for processing input signals in the first encoding means and the means for processing input signals  
20     in the second encoding means are each arranged to define groups containing at least three atoms.

5.     A system according to any preceding claim, wherein the means for processing input signals in the first  
25     encoding means and the means for processing input signals in the second encoding means are each arranged to process input signals defining a molecule in terms of a Molfile.

- 38 -

6. A system according to any preceding claim, wherein the means for processing input signals in the first encoding means and the means for processing input signals in the second encoding means are each arranged to  
5 identify atoms within an input molecule which satisfy at least one predetermined criterion, and to define the groups from the identified atoms.

7. A system according to claim 6, wherein the means for  
10 processing input signals in the first encoding means and the means for processing input signals in the second encoding means are each arranged to identify active atoms which are likely to cause the input molecule to react, and to define the groups from the identified active  
15 atoms.

8. A system according to any preceding claim, wherein the means for processing input signals in the first encoding means and the means for processing input signals  
20 in the second encoding means are each arranged to define a group of atoms using information identifying the atoms within the group and information relating to the relative positions of the atoms within the group.

25 9. A system according to any preceding claim, wherein:  
the means for processing input signals in the first encoding means and the means for processing input signals

- 39 -

in the second encoding means are each arranged to define at least one physical property of each input molecule; and

the means for generating signals in the first  
5 encoding means and the means for generating signals in the second encoding means are each arranged to generate the signals defining the molecule using further the defined physical properties of the molecule.

10 10. A system according to claim 9, wherein the means for processing input signals in the first encoding means and the means for processing input signals in the second encoding means are each arranged to define a physical  
15 property comprising the relative amount of atoms within the molecule which are of a predetermined type.

11. A system according to any preceding claim, wherein in the first encoding means:

the means for processing the defined groups and the  
20 input signals defining a property of each of the plurality of molecules is arranged to define the characteristic of each group by calculating a value for the group for each molecule property; and

the means for generating signals is arranged to  
25 define each molecule in dependence upon the cumulative values for each group in the molecule.

- 40 -

12. A system according to claim 11, wherein the means for generating signals is arranged to define each molecule in dependence upon a histogram of the values for each group in the molecule.

5

13. A system according to any preceding claim, wherein the means for generating signals in the first encoding means and the means for generating signals in the second encoding means are each arranged to generate signals of  
10 a fixed length defining the molecule.

14. A method according to any preceding claim, wherein the predicting means comprises a neural network.

15 15. A method according to any of claims 1 to 13, wherein the predicting means comprises means arranged to perform statistical analysis.

16. A method of processing signals defining input  
20 molecules to produce signals conveying a predicted property of a molecule, comprising the steps of:

(a) first encoding, comprising:

processing input signals defining a plurality of molecules to define groups of atoms within each molecule;

25 processing the defined groups and input signals defining a property of each of the plurality of molecules to define a characteristic of each group in dependence

- 41 -

upon the property of molecules containing the group; and  
generating signals defining each of the plurality  
of molecules in dependence upon the defined  
characteristics;

5 (b) second encoding, comprising:

processing input signals defining a further molecule  
to define groups of atoms within the further molecule;  
and

generating signals defining the further molecule  
10 using the defined groups in the further molecule and the  
characteristics of the groups defined by the first  
encoding step; and

(c) generating a signal defining a predicted  
property of the further molecule by processing the  
15 signals generated in the first encoding step defining  
each of the plurality of molecules, signals defining the  
property of each of the plurality of molecules, and the  
signals generated in the second encoding step defining  
the further molecule.

20

17. A method according to claim 16, wherein, in the step  
of processing input signals in the first encoding step  
and in the step of processing input signals in the second  
encoding step, groups containing a predetermined number  
25 of atoms are defined.

18. A method according to claim 17, wherein, in the step

- 42 -

of processing input signals in the first encoding step and in the step of processing input signals in the second encoding step, all unique groups which contain the predetermined number of atoms are defined.

5

19. A method according to any of claims 16 to 18, wherein, in the step of processing input signals in the first encoding step and in the step of processing input signals in the second encoding step, groups containing  
10 at least three atoms are defined.

20. A method according to any of claims 16 to 19, wherein, in the step of processing input signals in the first encoding step and in the step of processing input  
15 signals in the second encoding step, input signals defining a molecule in terms of a Molfile are processed.

21. A method according to any of claims 16 to 20, wherein, in the step of processing input signals in the  
20 first encoding step and in the step of processing input signals in the second encoding step, atoms are identified within an input molecule which satisfy at least one predetermined criterion, and the groups are defined from the identified atoms.

25

22. A method according to claim 21, wherein, in the step of processing input signals in the first encoding step

- 43 -

and in the step of processing input signals in the second encoding step, active atoms which are likely to cause the input molecule to react are identified, and the groups are defined from the identified active atoms.

5

23. A method according to any of claims 16 to 22, wherein, in the step of processing input signals in the first encoding step and in the step of processing input signals in the second encoding step, a group of atoms is  
10 defined using information identifying the atoms within the group and information relating to the relative positions of the atoms within the group.

24. A method according to any of claims 16 to 23,  
15 wherein:

in the step of processing input signals in the first encoding step and in the step of processing input signals in the second encoding step, at least one physical property of each input molecule is also defined; and

20 in the step of generating signals in the first encoding step and in the step of generating signals in the second encoding step, the signals defining the molecule are generated using the defined physical properties of the molecule as well.

25

25. A method according to claim 24, wherein, in the step of processing input signals in the first encoding step

- 44 -

and in the step of processing input signals in the second encoding step, a physical property is defined comprising the relative amount of atoms within the molecule which are of a predetermined type.

5

26. A system according to any of claims 16 to 25, wherein, in the first encoding step:

in the step of processing the defined groups and the input signals defining a property of each of the plurality of molecules, the characteristic of each group is defined by calculating a value for the group for each molecule property; and

in the step of generating signals, each molecule is defined in dependence upon the cumulative values for each group in the molecule.

27. A method according to claim 26, wherein, in the step of generating signals, each molecule is defined in dependence upon a histogram of the values for each group in the molecule.

28. A method according to any of claims 16 to 27, wherein, in the step of generating signals in the first encoding step and in the step of generating signals in the second encoding step, signals of a fixed length defining the molecule are generated.

- 45 -

29. A method according to any of claims 16 to 28,  
wherein a neural network is used in the predicting step.

30. A method according to any of claims 16 to 28,  
5 wherein statistical analysis is used in the predicting  
step.

31. A signal processing system, comprising:

(a) encoding means comprising:

10 storage means storing data defining a plurality of  
groups of atoms and a characteristic associated with each  
group;

means for processing input signals defining a  
molecule to define groups of atoms within the molecule;

15 and

means for generating signals defining the molecule  
using the defined groups and the stored data; and

(b) trained predicting means, comprising predicting  
means trained on the basis of data corresponding to the  
20 data stored in the encoding means for processing the  
signals generated by the encoding means defining the  
molecule to generate a signal defining a predicted  
property of the molecule.

25 32. Signal processing apparatus, comprising:

(a) first encoding means, comprising:

means for processing input signals defining a

- 46 -

plurality of molecules to define groups of atoms within each molecule;

means for processing the defined groups and input signals defining a property of each of the plurality of molecules to define a characteristic of each group in dependence upon the property of molecules containing the group; and

means for generating signals defining each of the plurality of molecules in dependence upon the defined characteristics; and

(b) second encoding means, comprising:

means for processing input signals defining a further molecule to define groups of atoms within the further molecule; and

means for generating signals defining the further molecule using the defined groups in the further molecule and the characteristics of the groups defined by the first encoding means.

33. Signal processing apparatus, comprising:

means for processing input signals defining a plurality of molecules to define groups of atoms within each molecule;

means for processing the defined groups and input signals defining a property of each of the plurality of molecules to define a characteristic of each group in dependence upon the property of molecules containing the

- 47 -

group; and

means for generating signals defining each of the plurality of molecules in dependence upon the defined characteristics.

5

34. Signal processing apparatus, comprising:

storage means storing data defining a plurality of groups of atoms and a characteristic associated with each group;

10 means for processing input signals defining a molecule to define groups of atoms within the molecule; and

means for generating signals defining the molecule using the defined groups and the stored data.

15

35. Signal processing apparatus, comprising:

means for processing input signals defining a molecule to define groups of atoms within the molecule.

20 36. Signal processing apparatus, comprising:

means for processing input signals defining groups of atoms within a plurality of molecules and signals defining a property of each of the plurality of molecules to generate signals defining a characteristic of each group in dependence upon the property of molecules containing the group.

25

- 48 -

37. Signal processing apparatus, comprising:

means for processing signals defining groups of atoms within a plurality of molecules, and signals defining a characteristic of each group to generate  
5 signals defining at least one molecule in dependence upon the characteristic of each group in the molecule.

38. Signal processing apparatus, comprising trained predicting means operable to process input signals  
10 defining an input molecule to generate a signal defining a predicted property of the molecule, wherein:

(i) the predicting means has been trained by:

processing input signals defining a plurality of molecules to define groups of atoms within each molecule;  
15 processing the defined groups and input signals defining a property of each of the plurality of molecules to define a characteristic of each group in dependence upon the property of molecules containing the group; and  
generating signals defining each of the plurality  
20 of molecules in dependence upon the defined characteristics; and

inputting the signals defining each of the plurality of molecules together with signals defining the property of each of the plurality of molecules into the predicting  
25 means; and

(ii) the input signals defining the input molecule comprise signals generated by processing signals defining

- 49 -

the molecule to define groups of atoms with the molecule, and processing the defined groups within the molecule and the characteristics of the groups defined during training to generate input signals defining the molecule.

5

39. A method of processing signals defining a molecule to produce signals conveying a predicted property of the molecule, comprising:

(a) an encoding step comprising:

10 processing input signals defining a molecule to define groups of atoms within the molecule; and

generating signals defining the molecule using the defined groups and encoding data defining a plurality of groups of atoms and a characteristic associated with each

15 group; and

(b) processing the signals generated by the encoding step using trained predicting means, comprising predicting means trained on the basis of data corresponding to the encoding data used in the encoding  
20 step, to generate a signal defining a predicted property of the molecule.

40. A method of processing signals to produce signals encoding a molecule, comprising the steps of:

25 (a) first encoding, comprising:

processing input signals defining a plurality of molecules to define groups of atoms within each molecule;

- 50 -

processing the defined groups and input signals  
defining a property of each of the plurality of molecules  
to define a characteristic of each group in dependence  
upon the property of molecules containing the group; and

5       generating signals defining each of the plurality  
of molecules in dependence upon the defined  
characteristics; and

(b) second encoding, comprising:

processing input signals defining a further molecule  
10   to define groups of atoms within the further molecule;  
and

generating signals defining the further molecule  
using the defined groups in the further molecule and the  
characteristics of the groups defined in the first  
15   encoding step.

41. A method of processing signals to produce signals  
encoding molecules, comprising the steps of:

processing input signals defining a plurality of  
20   molecules to define groups of atoms within each molecule;

processing the defined groups and input signals  
defining a property of each of the plurality of molecules  
to define a characteristic of each group in dependence  
upon the property of molecules containing the group; and

25       generating signals defining each of the plurality  
of molecules in dependence upon the defined  
characteristics.

- 51 -

42. A method of processing signals to produce signals encoding molecules, comprising the steps of:

processing input signals defining a molecule to define groups of atoms within the molecule; and

5 generating signals defining the molecule using the defined groups and data defining a plurality of groups of atoms and a characteristic associated with each group.

43. A signal processing method, comprising processing  
10 input signals defining a molecule to define groups of atoms within the molecule.

44. A signal processing method, comprising processing  
input signals defining groups of atoms within a plurality  
15 of molecules and signals defining a property of each of the plurality of molecules to generate signals defining a characteristic of each group in dependence upon the property of molecules containing the group.

20 45. A method of processing signals to produce signals encoding a molecule, comprising processing signals defining groups of atoms within a plurality of molecules, and signals defining a characteristic of each group to generate signals defining at least one molecule in  
25 dependence upon the characteristic of each group in the molecule.

- 52 -

46. A signal processing method, comprising processing input signals defining an input molecule using trained predicting means to generate a signal defining a predicted property of the molecule, wherein:

5 (i) the predicting means has been trained by:

processing input signals defining a plurality of molecules to define groups of atoms within each molecule;

processing the defined groups and input signals defining a property of each of the plurality of molecules  
10 to define a characteristic of each group in dependence upon the property of molecules containing the group; and

generating signals defining each of the plurality of molecules in dependence upon the defined characteristics; and

15 inputting the signals defining each of the plurality of molecules together with signals defining the property of each of the plurality of molecules into the predicting means; and

(ii) the input signals defining the input molecule  
20 comprise signals generated by processing signals defining the molecule to define groups of atoms within the molecule, and processing the defined groups within the molecule and the characteristics of the groups defined during training to generate input signals defining the molecule.

25

47. A method of processing signals to produce signals defining a molecule, comprising:

- 53 -

(a) performing a method according to claim 16, claim 39 or claim 46 to produce signals conveying a predicted property of a molecule;

(b) using the predicted property to determine  
5 whether to select the molecule; and

(c) if the molecule is selected in step (b), generating a signal defining the molecule.

48. A method according to any of claims 16 to 30 or 39  
10 to 47, further comprising the step of recording the signals either directly or indirectly.

49. A storage device storing instructions for causing a programmable processing apparatus to perform a method  
15 according to any of claims 16 to 30 or 39 to 48.

50. A signal conveying instructions for causing a programmable processing apparatus to perform a method according to any of claims 16 to 30 or 39 to 48.

20

51. A process of producing a compound, comprising:

(a) performing a method according to claim 16, claim 39 or claim 46 to predict a property of a molecule;

(b) using the predicted property to determine  
25 whether to select the molecule; and

(c) if the molecule is selected in step (b), producing a compound containing the molecule.

- 54 -

52. A process of manufacturing a compound, comprising:

(a) performing a method according to claim 16, claim 39 or claim 46 to predict a property for a plurality of molecules;

5 (b) selecting molecules for testing in dependence upon the predicted properties;

(c) testing the selected molecules; and

(d) manufacturing a compound containing at least one tested molecule.

10

53. A compound produced by:

(a) performing a method according to claim 16, claim 39 or claim 46 to predict a property of a molecule; and

15 (b) producing a compound containing the molecule.

54. A compound containing a molecule having a property predicted by a method according to claim 16, claim 39 or claim 46.

20

55. Use of a method according to claim 16, claim 39, claim 46 or claim 47 in the production of a compound.

56. Use of a molecule identified using a method  
25 according to claim 16, claim 39 or claim 46 in the production of a compound.

- 55 -

57. Use of a molecule defined by signals produced using a method according to claim 47 in the production of a compound.

1/22

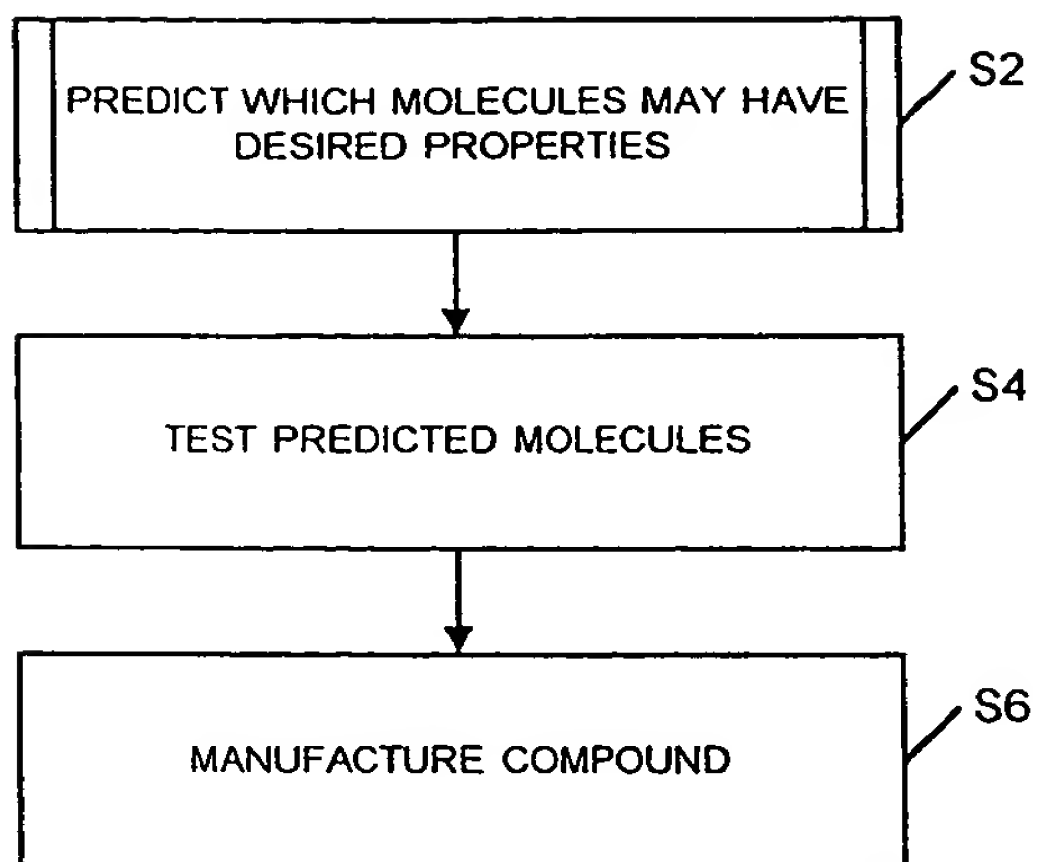


FIG. 1

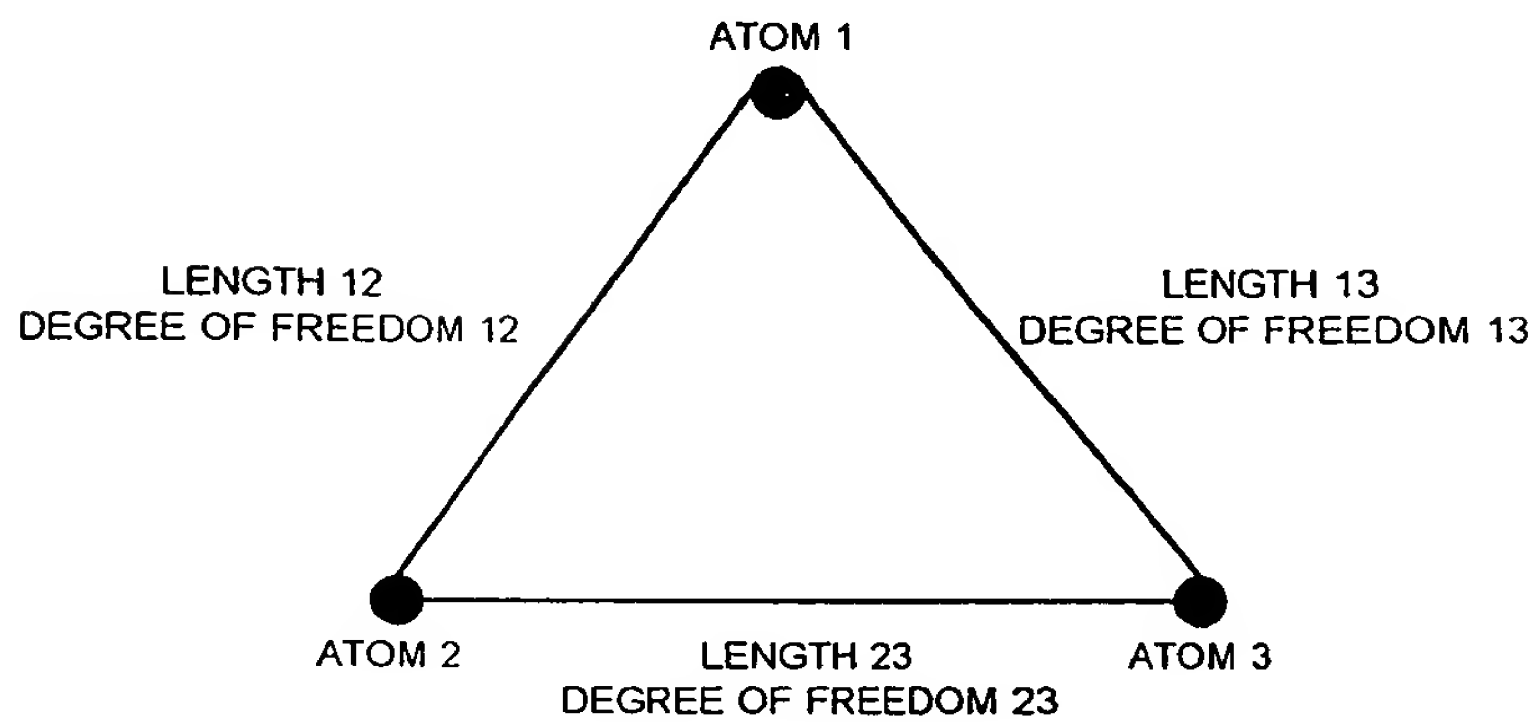


FIG. 7

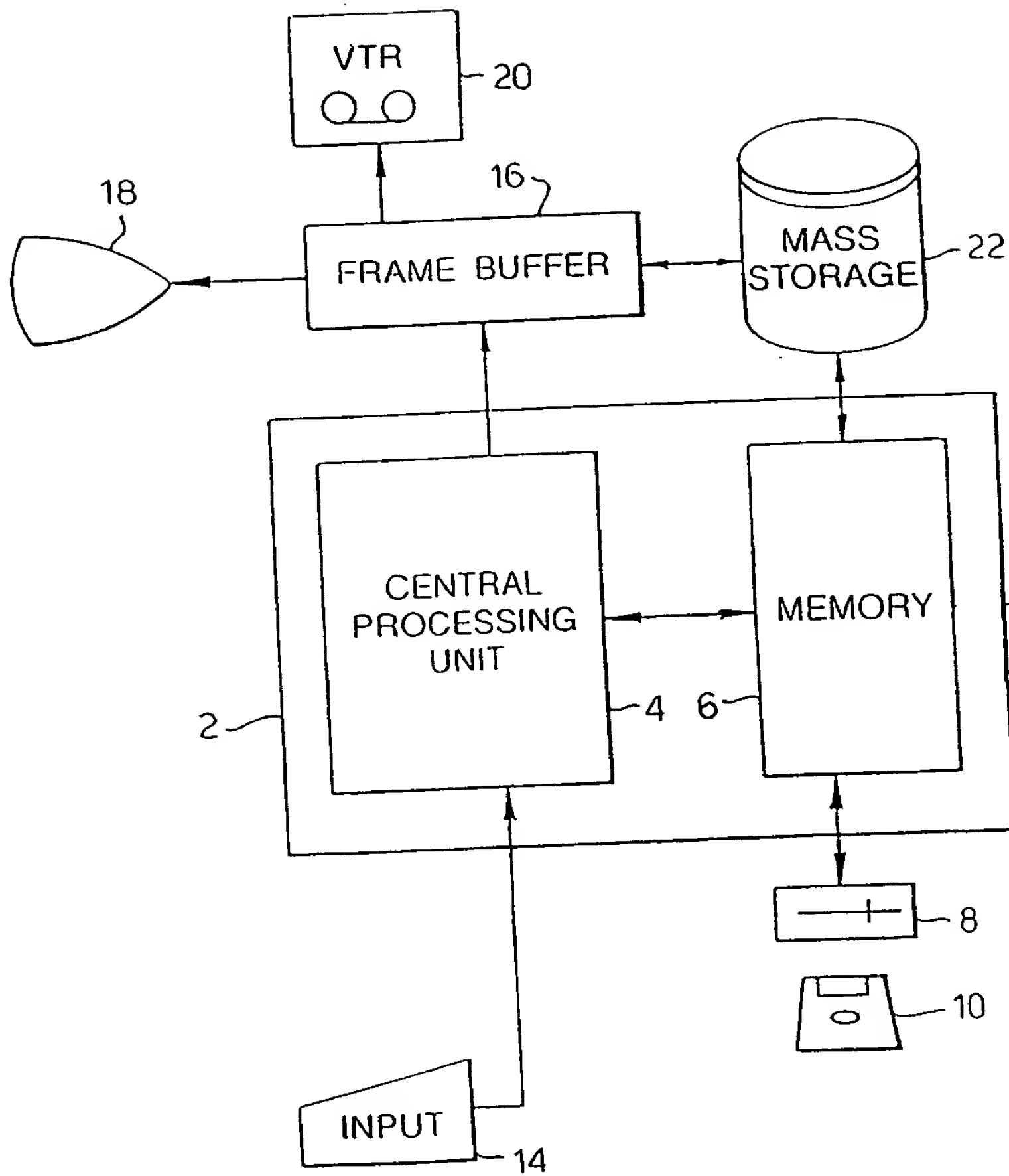


FIG. 2

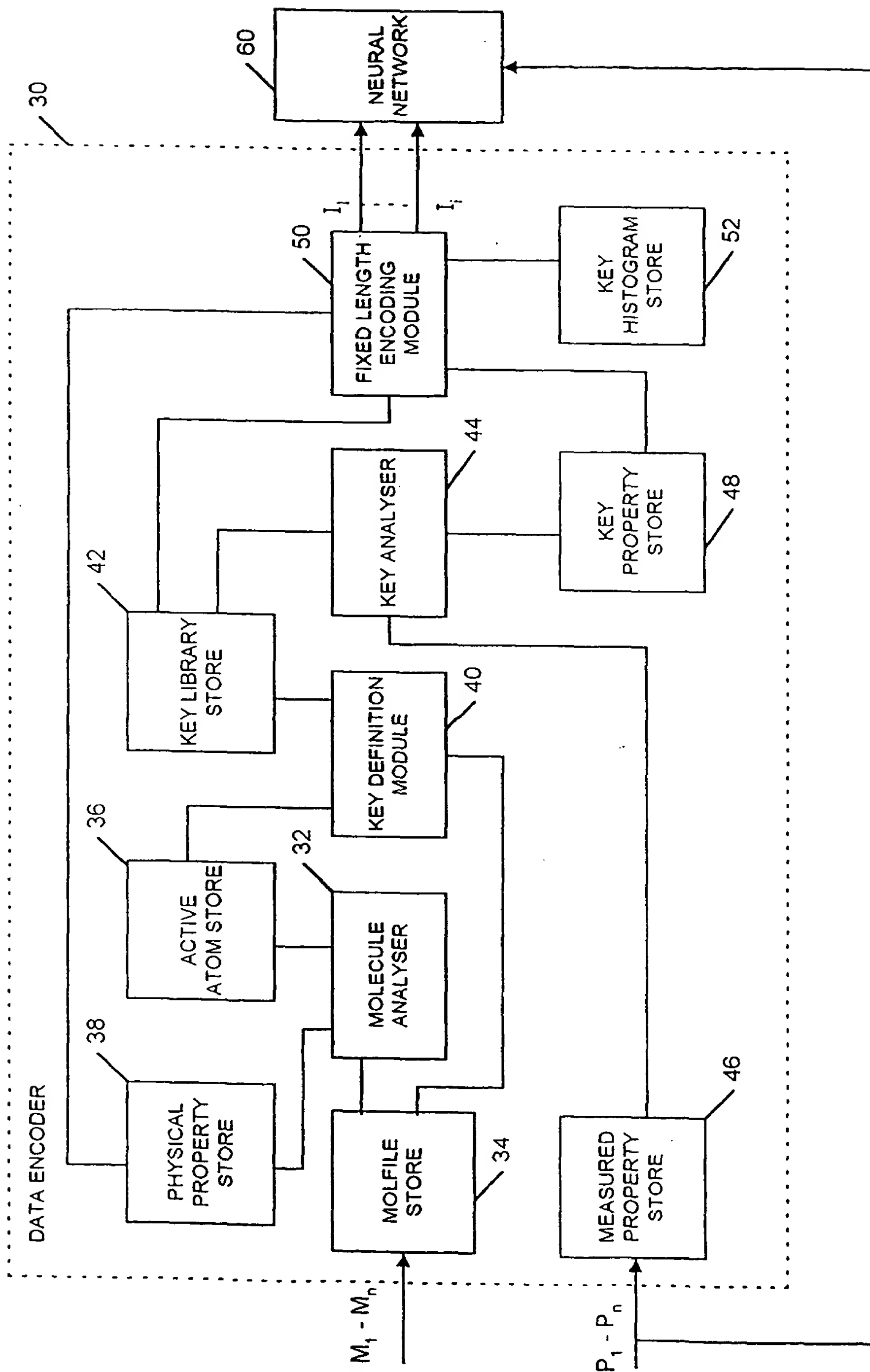


FIG. 3

4/22

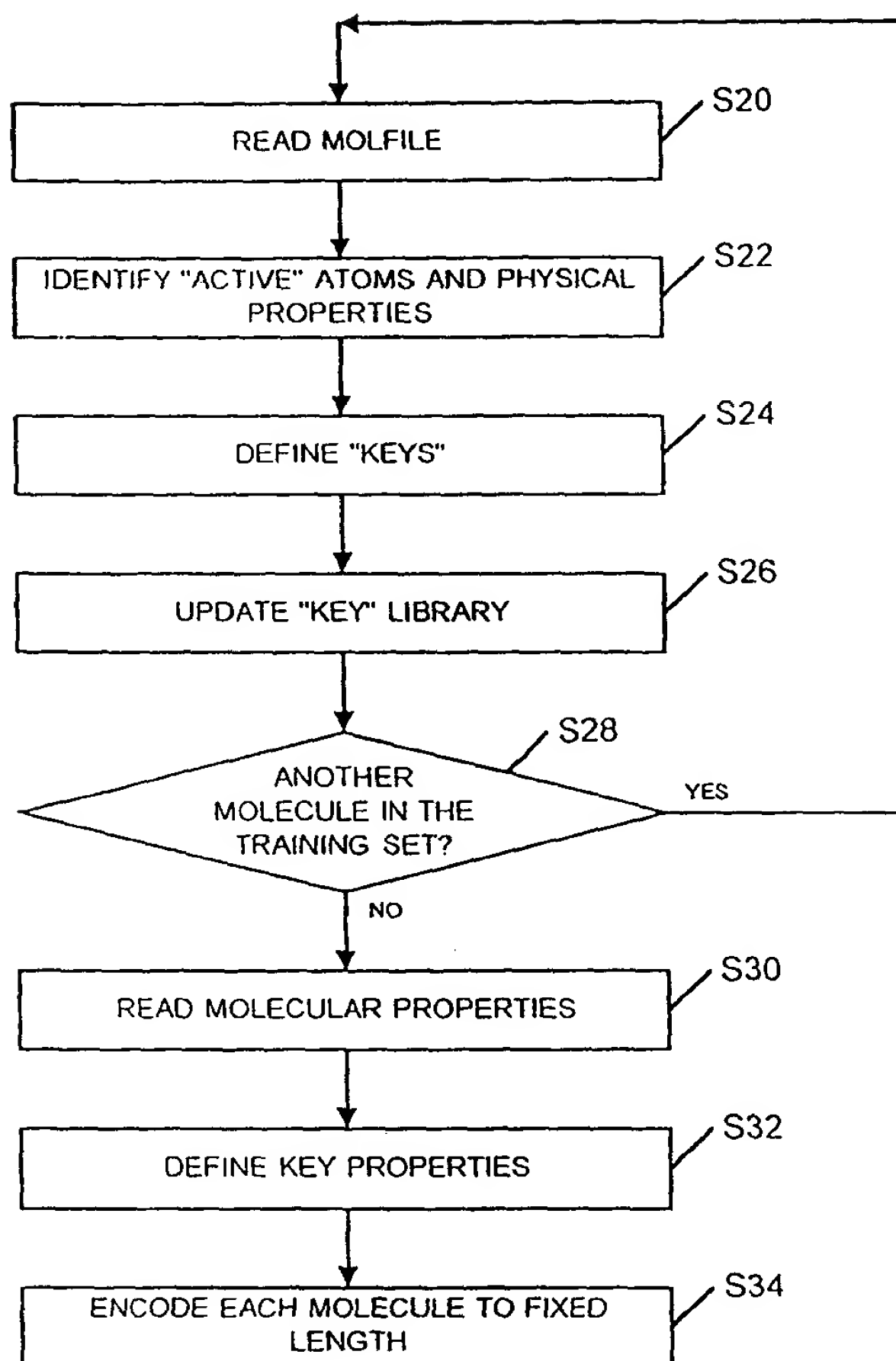


FIG. 4

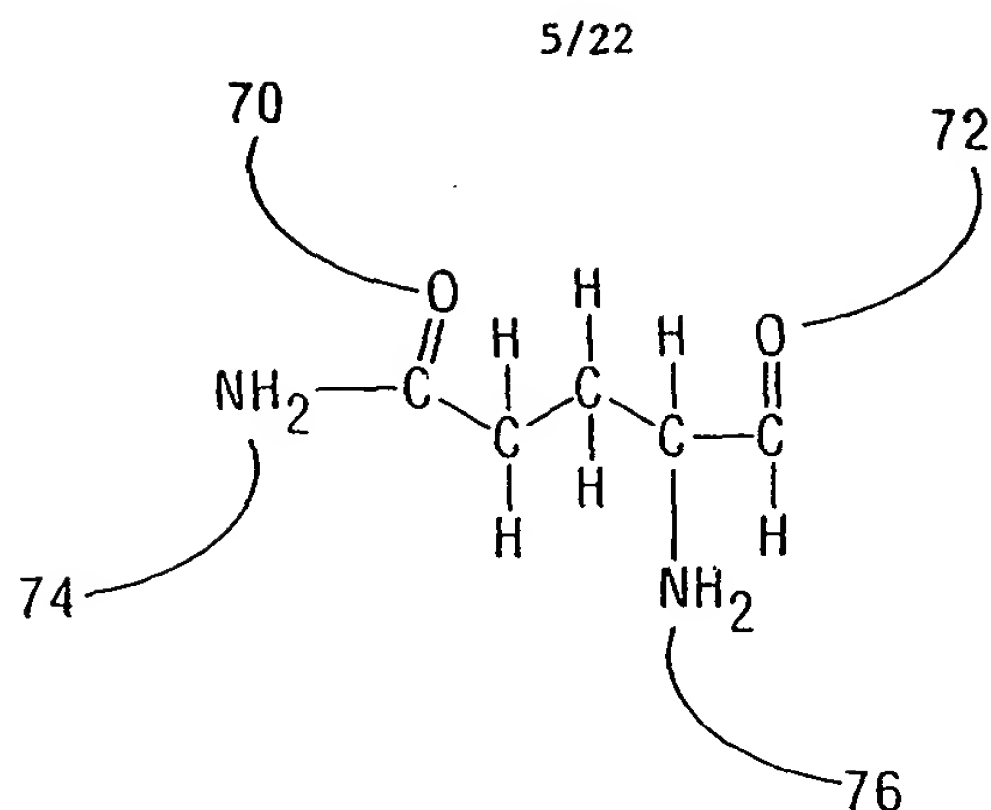


FIG. 5

-ISIS- 11209709492D

9 8 0 0 0 0 0 0 0 0 1 V2000

0.4949	-5.8251	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0
0.5070	-5.1063	0.0000	C	0	0	3	0	0	0	0	0	0	0	0	0	0
1.2123	-4.8573	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2123	-4.1102	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.1152	-4.6836	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.8209	-4.9399	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.2940	-4.3547	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.0336	-3.6332	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0
-2.0373	-4.4833	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0

2	1	1	0	0	0	0
2	3	1	0	0	0	0
3	4	2	0	0	0	0
2	5	1	0	0	0	0
6	5	1	0	0	0	0
7	8	2	0	0	0	0
6	7	1	0	0	0	0
7	9	1	0	0	0	0

M END

 HEADER BLOCK

 COUNTS LINE

 ATOM BLOCK

 BOND BLOCK

FIG. 6

6/22

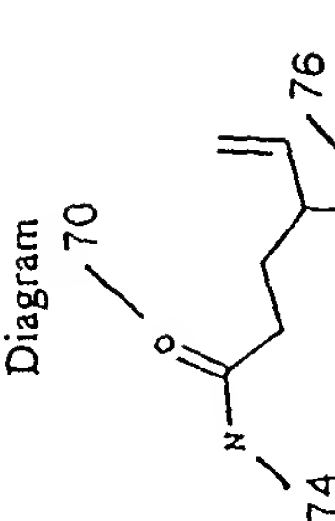
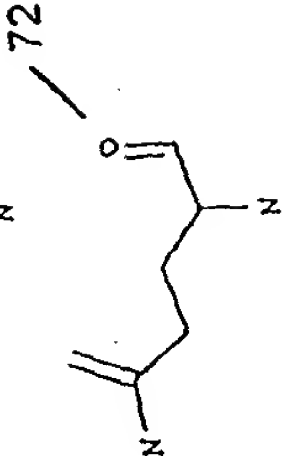
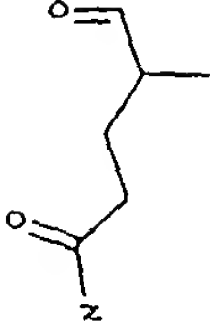
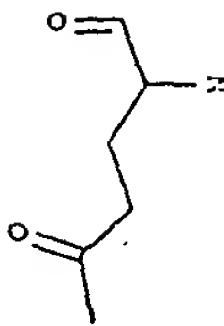
Atomic Numbers		Network Distances			Degrees of Freedom		Diagram		
7	7	8	5	2	5	4	1		
7	7	8	5	3	6	5	2	5	
7	8	8	6	6	2	5	4	1	
7	8	8	3	6	5	2	4	4	

FIG. 8

7/22

	MOLECULE 1
KEY 1	✓
KEY 2	✓
KEY 3	✓
KEY 4	✓

FIG. 9

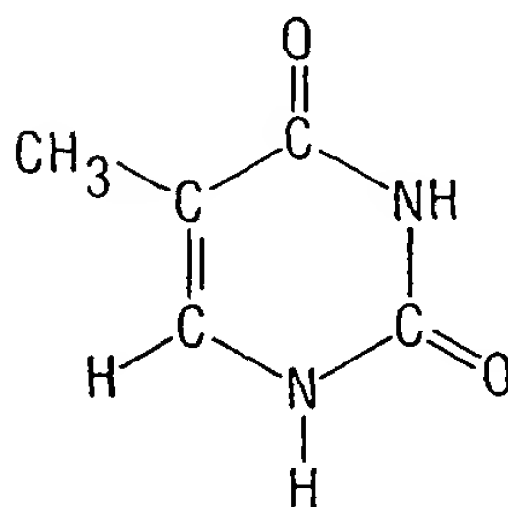


FIG. 10a

8/22

-ISIS- 11209710572D

9 9 0 0 0 0 0 0 0 0 0 1 V2000

7.9271	0.0588	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0
8.5732	0.4358	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
7.2730	0.4360	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
8.5732	1.1915	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0
9.2229	0.0629	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0
7.2731	1.1874	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
7.9239	1.5607	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0
7.9239	2.3120	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0
6.6271	1.5621	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0

1 2 1 0 0 0 0

1 3 1 0 0 0 0

2 4 1 0 0 0 0

2 5 2 0 0 0 0

3 6 2 0 0 0 0

4 7 1 0 0 0 0

7 8 2 0 0 0 0

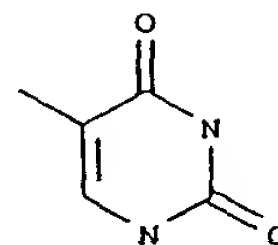
6 7 1 0 0 0 0

6 9 1 0 0 0 0

M END

FIG. 10b

Atom is not Hydrogen or  
Carbon



Atom is charged

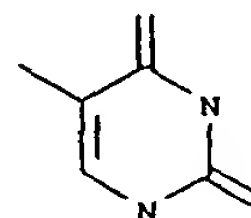


FIG. 10c

9/22

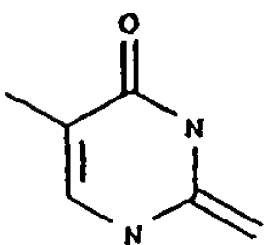
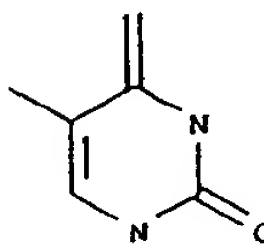
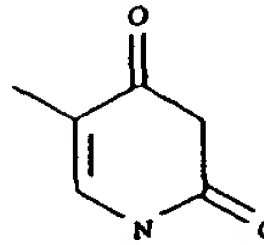
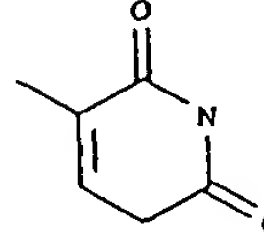
Atomic Numbers			Network Distances			Degrees of Freedom			Diagram
7	7	8	2	2	4	0	0	0	
7	7	8	2	2	2	0	0	0	
7	8	8	2	4	4	0	0	0	
7	8	8	2	4	2	0	0	0	

FIG. 10d

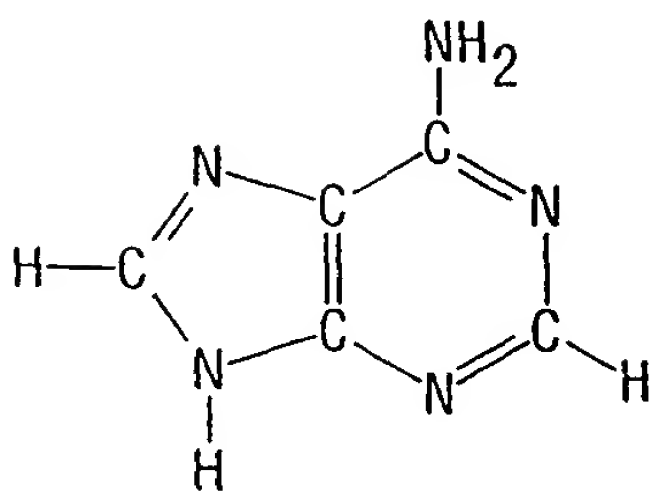


FIG. 11a

10/22

-ISIS- 11209710332D

10 11 0 0 0 0 0 0 0 0 1 V2000

-1.6496	0.3794	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0
-1.0049	0.5975	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0
-2.0940	1.0249	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0
-1.0067	1.4090	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0
-0.3620	0.2136	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0
-1.6443	1.6369	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0
-0.3815	1.7750	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0
0.3018	0.5896	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0
0.3065	1.3783	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0
-0.3823	2.4906	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0

1 2 1 0 0 0 0

1 3 1 0 0 0 0

2 4 2 0 0 0 0

2 5 1 0 0 0 0

3 6 2 0 0 0 0

4 7 1 0 0 0 0

5 8 2 0 0 0 0

7 9 2 0 0 0 0

7 10 1 0 0 0 0

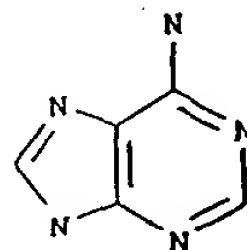
4 6 1 0 0 0 0

8 9 1 0 0 0 0

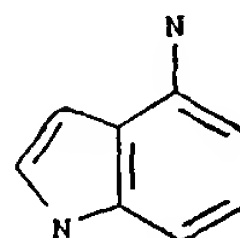
M END

FIG. 11b

Atom is not Hydrogen or  
Carbon



Atom is charged



Active atom being a  
virtual atom at the centre  
of an aromatic ring

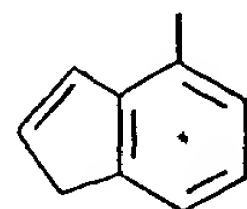


FIG. 11c

SUBSTITUTE SHEET (RULE 26)

11/22

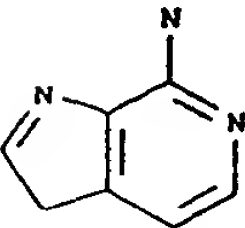
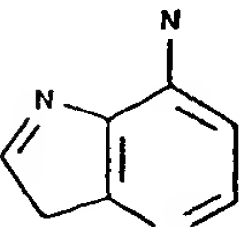
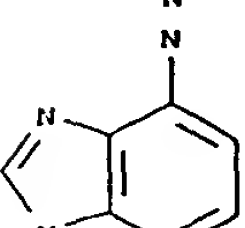
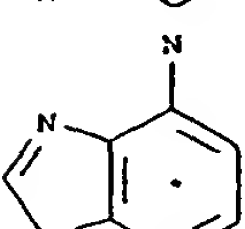
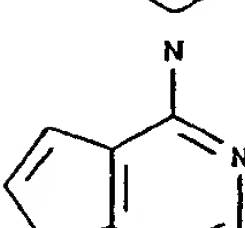
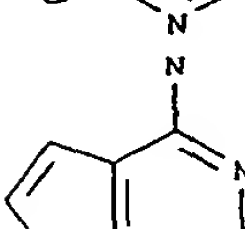
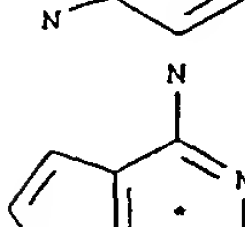
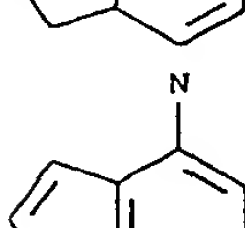
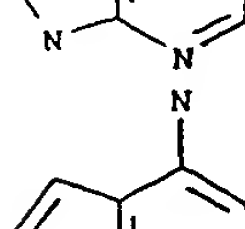
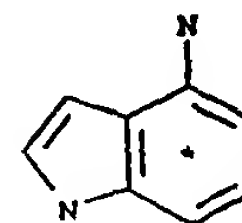
Atomic Numbers			Network Distances			Degrees of Freedom			Diagram
7	7	7	3	3	2	1	0	1	
7	7	7	3	3	4	1	0	1	
7	7	7	3	2	4	1	0	1	
7	7	0	3	2	2	1	0	1	
7	7	7	2	2	4	1	0	1	
7	7	7	2	4	4	1	0	1	
7	7	0	2	1	2	1	0	1	
7	7	7	4	2	4	1	0	1	
7	7	0	4	1	2	1	0	1	

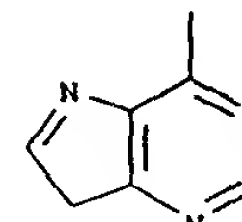
FIG. 11d

12/22

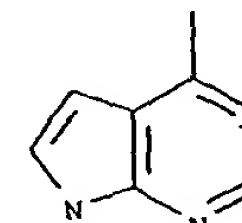
7 7 0 4 2 2 1 0 1



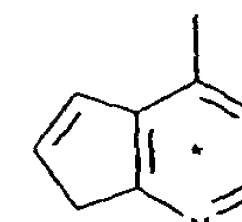
7 7 7 2 3 3 0 0 0



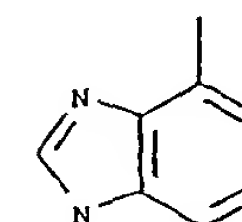
7 7 7 2 2 4 0 0 0



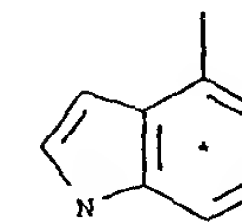
7 7 0 2 1 1 0 0 0



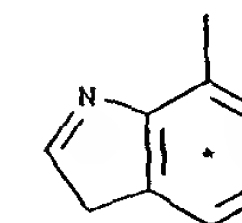
7 7 7 4 2 3 0 0 0



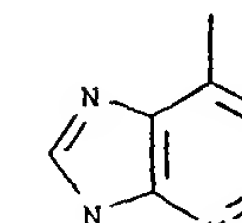
7 7 0 4 2 1 0 0 0



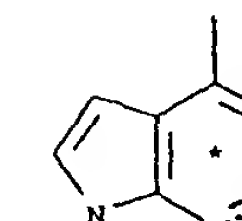
7 7 0 3 2 1 0 0 0\*



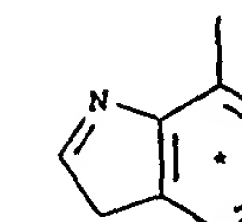
7 7 7 2 2 3 0 0 0



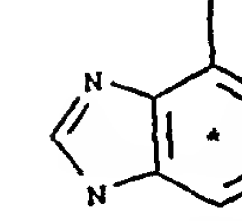
7 7 0 2 2 1 0 0 0



7 7 0 3 2 1 0 0 0\*



7 7 0 2 2 2 0 0 0

FIG. 11d (Cont)  
SUBSTITUTE SHEET (RULE 26)

13/22

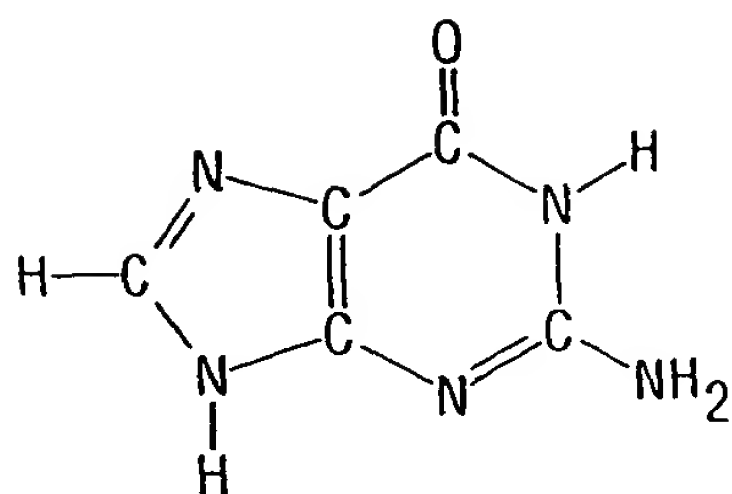


FIG. 12a

-ISIS- 11209710392D

```

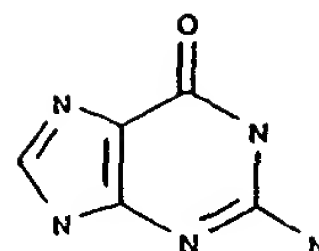
11 12 0 0 0 0 0 0 0 0 0 1 V2000
  2.5357 0.2845 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  3.1767 0.5019 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  2.0877 0.9308 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  3.1758 1.3180 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  3.8234 0.1173 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  2.5423 1.5424 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  3.8056 1.6833 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  4.4879 0.4969 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  4.4936 1.2860 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  3.8056 2.4035 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  5.1165 0.1306 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0 0 0 0
1 3 1 0 0 0 0
2 4 2 0 0 0 0
2 5 1 0 0 0 0
3 6 2 0 0 0 0
4 7 1 0 0 0 0
5 8 2 0 0 0 0
7 9 1 0 0 0 0
7 10 2 0 0 0 0
4 6 1 0 0 0 0
8 9 1 0 0 0 0
8 11 1 0 0 0 0
|M END

```

FIG. 12b

SUBSTITUTE SHEET (RULE 26)

14/22

Atom is not Hydrogen or  
Carbon

Atom is charged

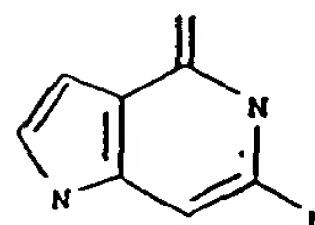


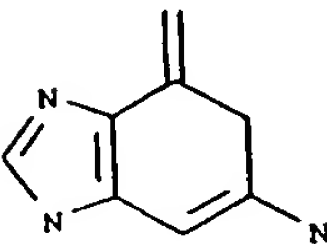
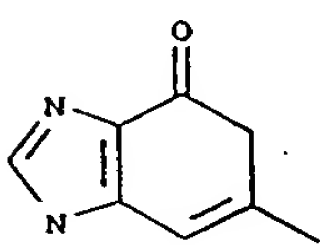
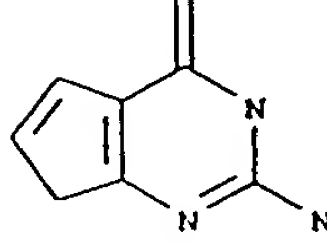
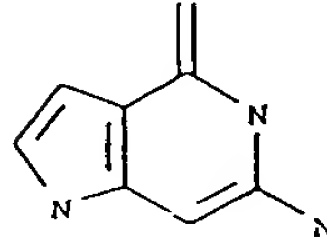
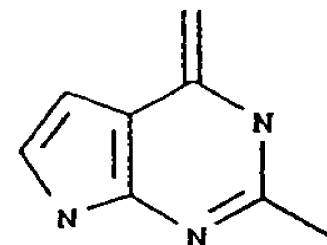
FIG. 12c

Atomic Numbers			Network Distances			Degrees of Freedom			Diagram
7	7	7	3	2	5	0	1	1*	
7	7	7	3	2	3	0	0	0	
7	7	7	3	4	2	0	0	0	
7	7	8	3	2	3	0	0	0	
7	7	7	5	2	3	1	1	0*	

FIG. 12d

SUBSTITUTE SHEET (RULE 26)

15/22

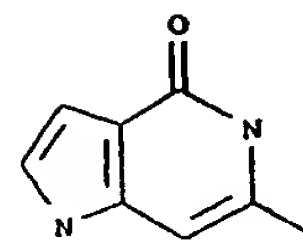
7	7	7	5	4	2	1	1	0	
7	7	8	5	4	3	1	1	0	
7	7	7	3	2	2	0	0	0	
7	7	8	3	4	3	0	0	0	
7	7	8	2	4	3	0	0	0	
7	7	7	2	2	2	1	1	0	
7	7	7	2	4	4	1	1	0	
7	7	8	2	4	2	1	1	0	
7	7	7	2	2	4	0	0	0	
7	7	8	2	4	2	0	0	0	

SUBSTITUTE SHEET (RULE 26)

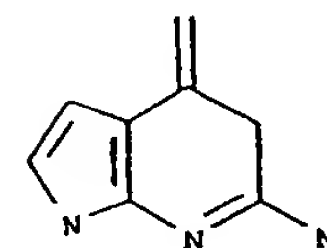
FIG. 12d (Cont)

16/22

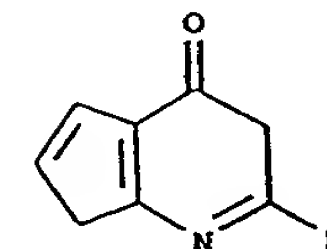
7 7 8 4 4 2 0 0 0



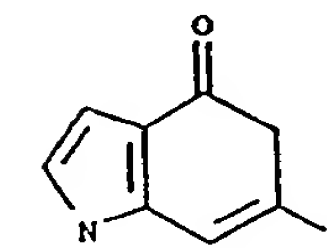
7 7 7 2 2 4 1 0 1



7 7 8 2 4 4 1 0 1



7 7 8 4 4 4 1 0 1



7 7 8 2 4 4 0 0 0

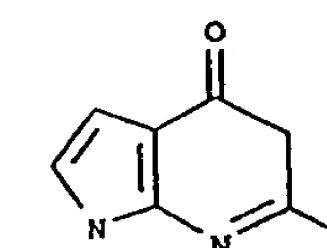


FIG. 12d (Cont)

	MOLECULE 1	MOLECULE 2	MOLECULE 3	MOLECULE 4
KEY 1	✓	—	✓	—
KEY 2	✓	—	—	✓
KEY 3	✓	✓	✓	—
KEY 4	✓	—	✓	—
KEY 5	—	✓	✓	—
KEY 6	—	✓	TWICE	✓
KEY 7	—	—	✓	—
KEY 8	—	—	—	TWICE
KEY 9	—	—	—	✓
KEY 10	—	—	—	✓

FIG. 13

18/22

	KEY PROPERTY				M4 MOLECULE 4	M3 MOLECULE 3	M2 MOLECULE 2	M1 MOLECULE 1	P1 0.5	P2 0.4	P3 0.8	P4 0.2
	KEY 1	KEY 2	KEY 3	KEY 4								
KEY 1	✓	—	✓	—	—	✓	—	✓	1/2	—	1/2	—
KEY 2	✓	—	—	✓	—	—	—	✓	1/2	—	—	1/2
KEY 3	✓	✓	✓	—	—	✓	—	✓	1/3	1/3	1/3	—
KEY 4	✓	—	✓	—	—	✓	—	✓	1/2	—	1/2	—
KEY 5	—	✓	✓	—	—	✓	—	—	—	1/2	1/2	—
KEY 6	—	✓	✓	✓	✓	✓	✓	—	—	1/4	1/2	1/4
KEY 7	—	—	✓	—	—	✓	—	—	—	—	1	—
KEY 8	—	—	—	✓	✓	—	—	—	—	—	—	1
KEY 9	—	—	—	✓	✓	—	—	—	—	—	—	1
KEY 10	—	—	—	✓	✓	—	—	—	—	—	—	1

FIG. 14

SUBSTITUTE SHEET (RULE 26)

19/22

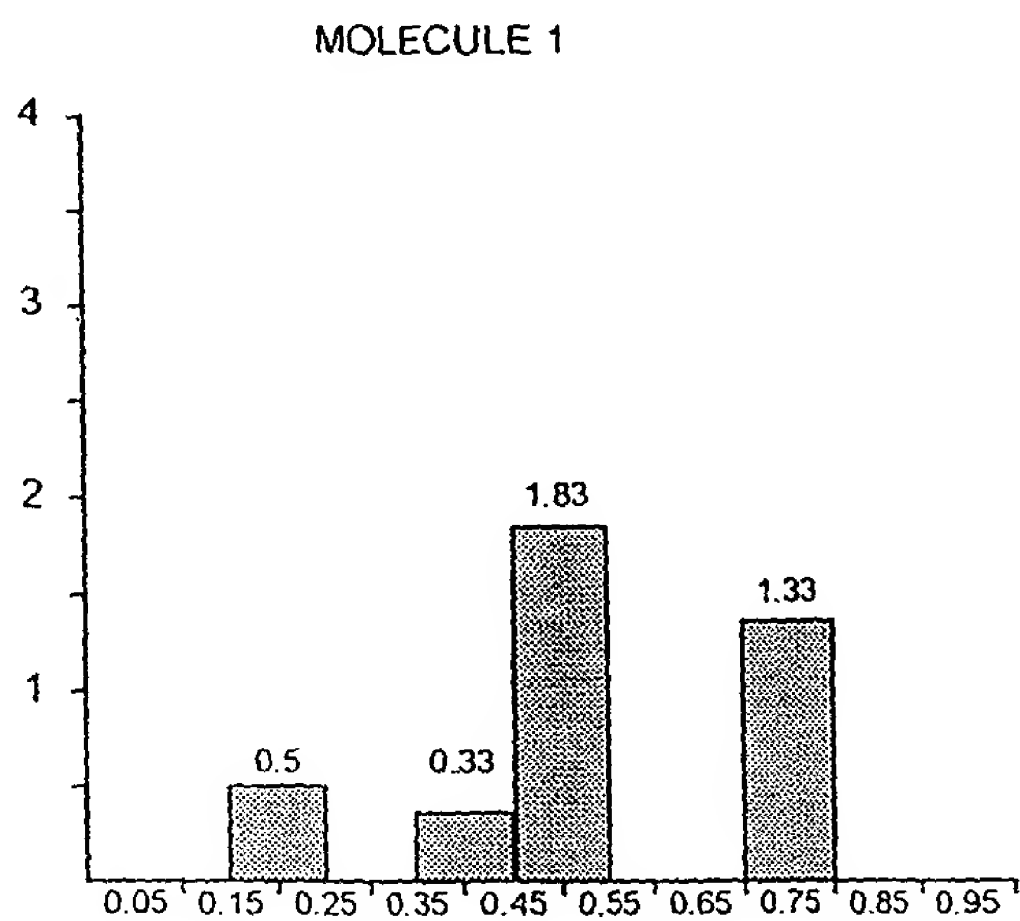


FIG. 15a

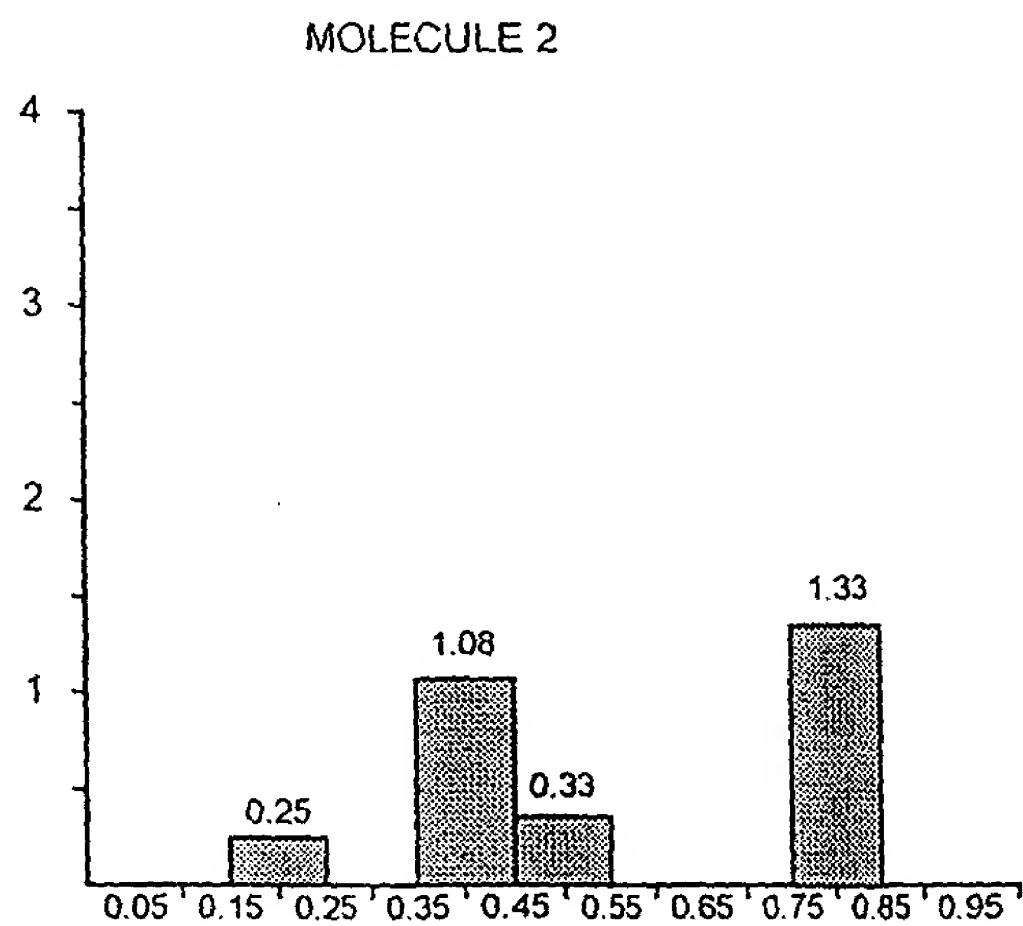


FIG. 15b

MOLECULE 3

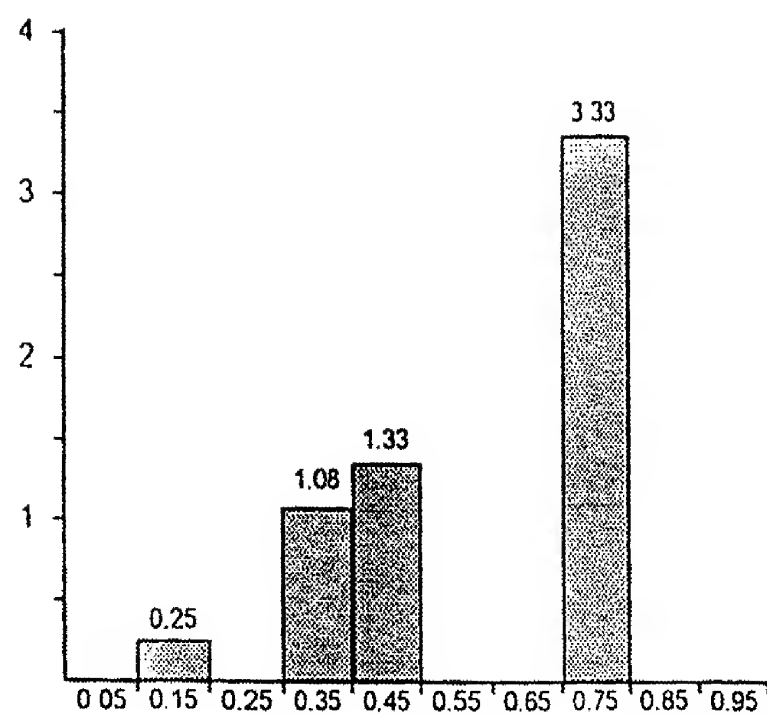


FIG. 15c

MOLECULE 4

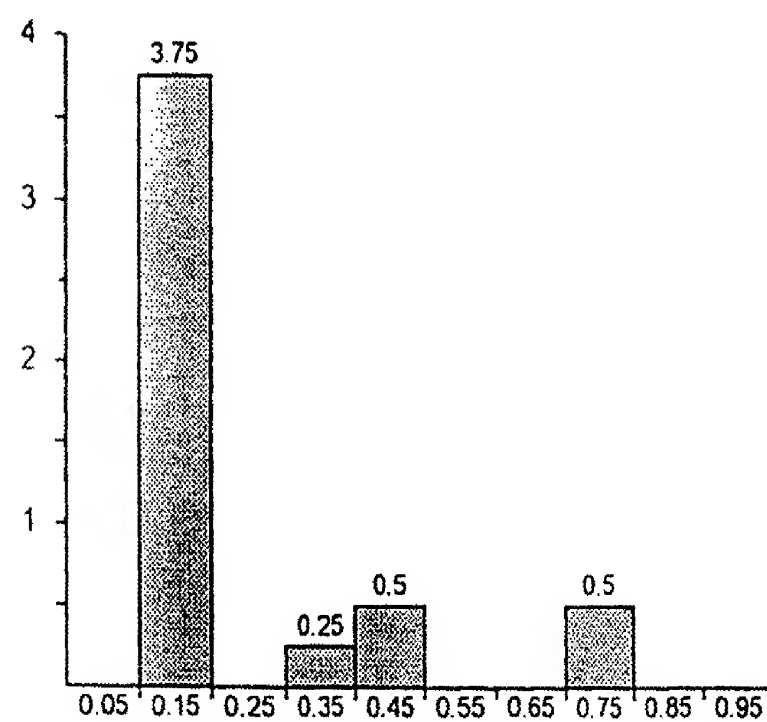


FIG. 15d

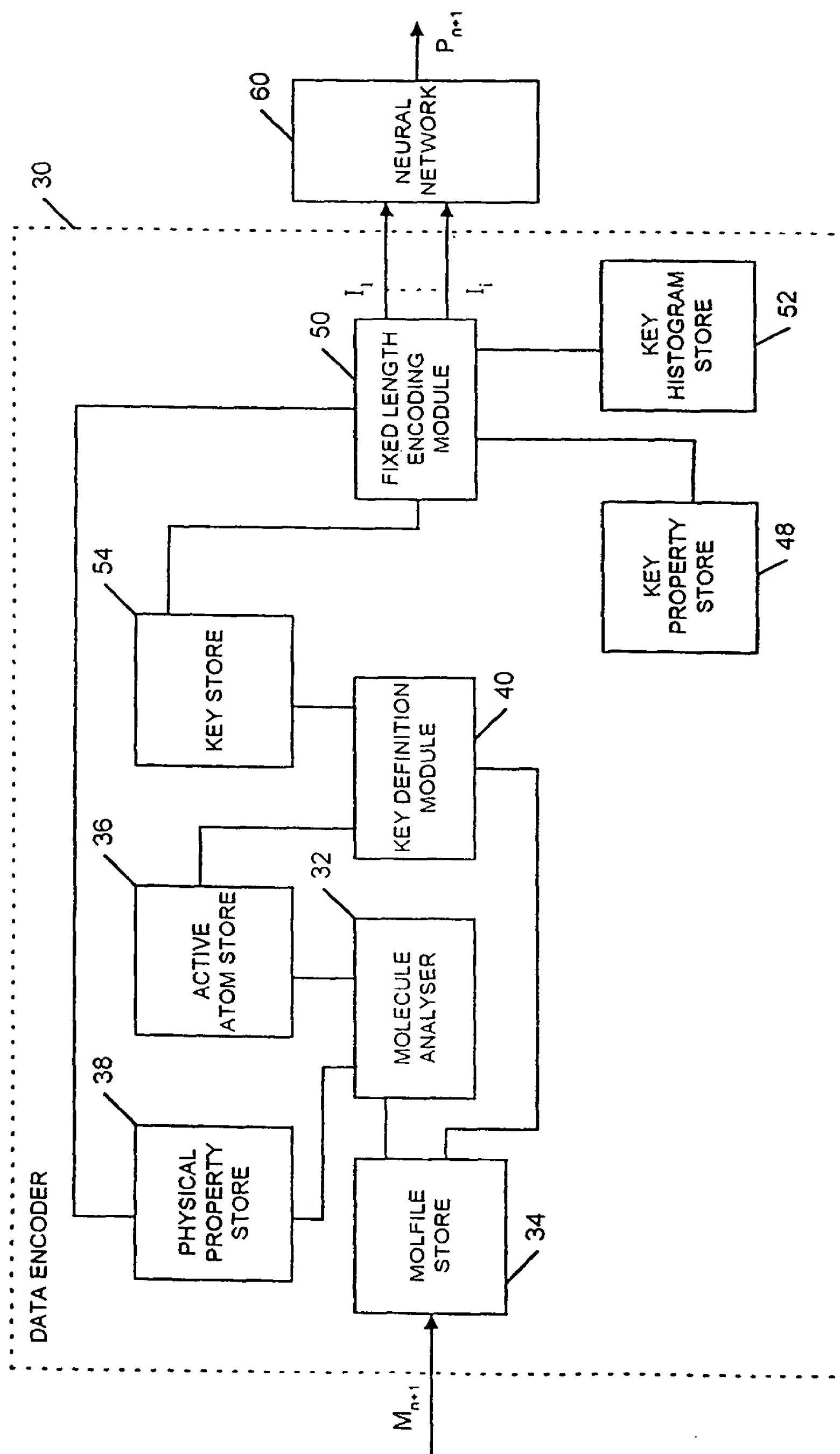


FIG. 16

22/22

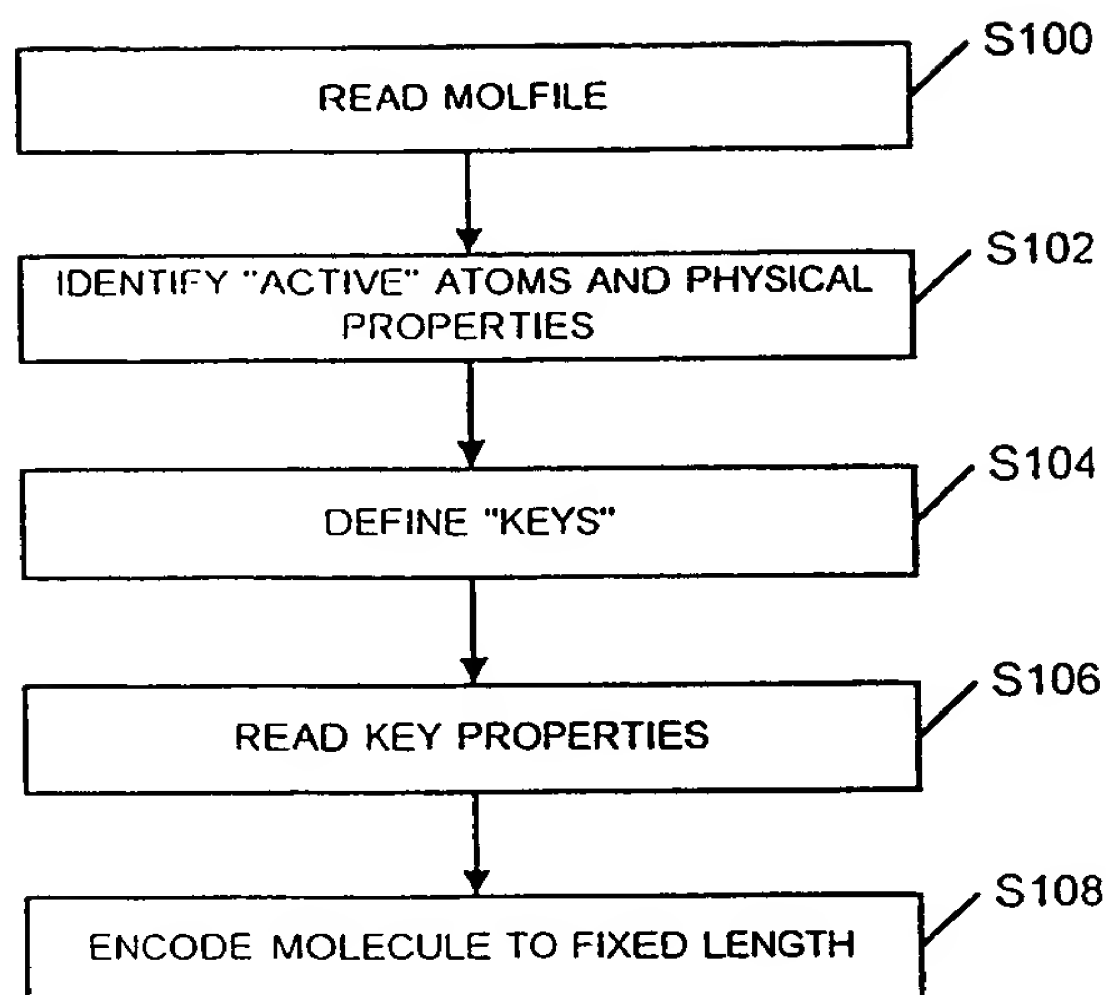


FIG. 17

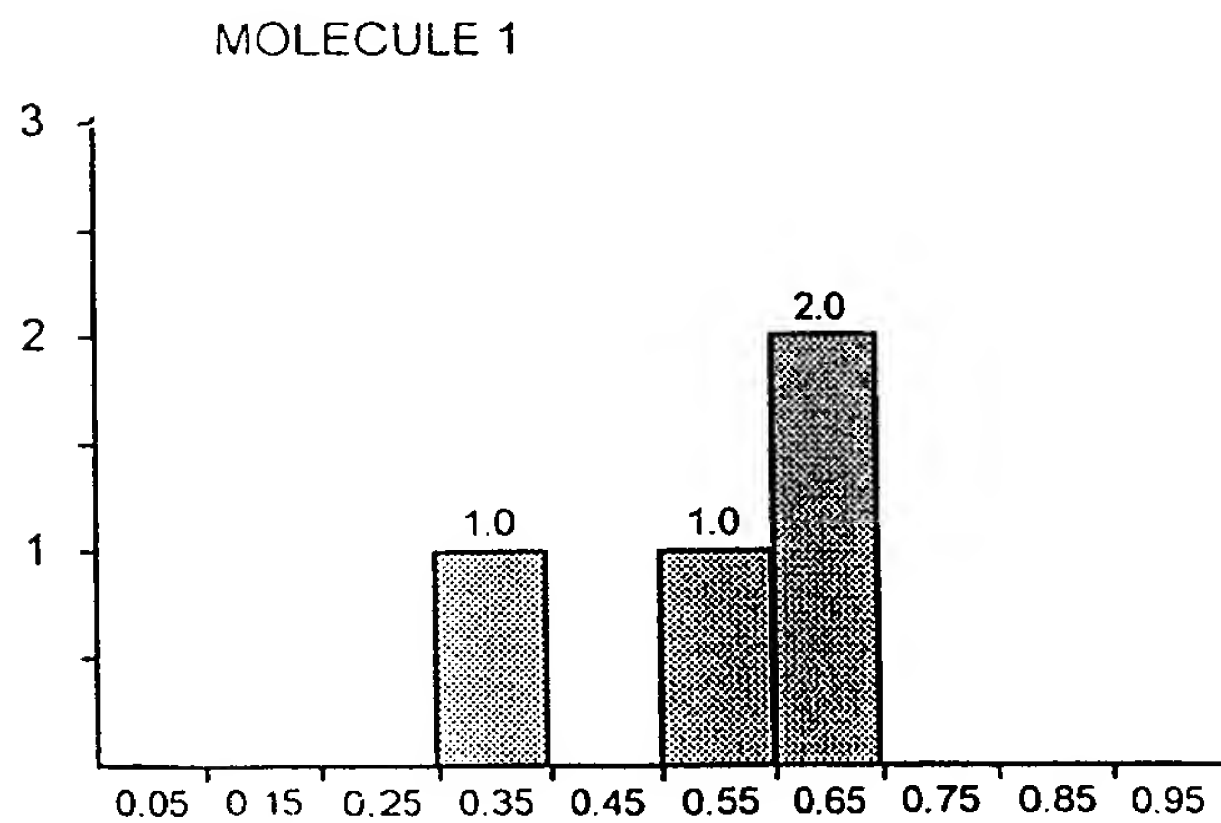


FIG. 18

# INTERNATIONAL SEARCH REPORT

International Application No  
PCT/GB 99/00046

**A. CLASSIFICATION OF SUBJECT MATTER**  
IPC 6 G06F17/50

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)  
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	EP 0 496 902 A (IBM ; IBM SEMEA (IT)) 5 August 1992 see page 3, line 11 - page 4, line 20; figures 1-3; tables 1-3  --- -/--	33-37, 41-45 1-10, 16-25, 31, 32, 38-40, 46-50

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

\* Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

21 April 1999

Date of mailing of the international search report

28/04/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Guingale, A